21-690: Methods of Optimization

Krish Matta

Contents

1 Introduction **2**

2 Convex Sets **2**
2.1 Affine and Convex Sets 2
2.2 Examples 4
2.2.1 Cones 4
2.2.2 Hyperplanes and Halfspaces 4
2.2.3 Balls and Ellipsoids 5
2.2.4 Polyhedra 6
2.3 Operations that Preserve Convexity 7
2.3.1 Intersection 7
2.3.2 Affine Functions 7
2.3.3 Perspective Function 8
2.4 Separating Hyperplanes 10
2.5 Supporting Hyperplanes 12

3 Convex Functions **13**
3.1 Basic Properties and Definitions 13
3.2 Operations that Preserve Convexity 16
3.2.1 Nonnegative Weighted Sum 16
3.2.2 Affine Composition 17
3.2.3 Maximum and Supremum 18
3.2.4 Representation as Supremum of Affine Functions . . . 18
3.2.5 Perspective of a Function 20
3.3 Convex Conjugates 20
3.3.1 Basic Properties 20

4 Convex Optimization Problems **21**
4.1 Optimization Problems 21

4.1.1 Slack Variables 22
4.2 Convex Problems 23
4.3 Linear Problems 26
4.4 Quadratic Problems 26

5 Duality **27**
5.1 Lagrange Dual Function 27
5.2 Lagrange Dual Problem 28
5.3 Geometric Intuition 29
5.3.1 Optimal Value 29
5.3.2 Lagrange Dual Function 30
5.3.3 Weak Duality 30
5.3.4 Epigraph Variation 31
5.3.5 Slater's Condition 31
5.4 Optimality Conditions 35
5.4.1 Certificate of Suboptimality 35
5.4.2 Complementary Slackness 35
5.4.3 KKT Conditions 36

6 Unconstrained Minimization **37**
6.1 Strong Convexity 37
6.1.1 Smoothness 40
6.2 Conditioning 41
6.3 Descent Methods 43
6.3.1 Exact Line Search 43
6.3.2 Backtracking 44
6.4 Gradient Descent 44

1 Introduction

Below are my notes for the course 21-690: Methods of Optimization taught in the Spring 2025 semester by Professor Nicholas Boffi at Carnegie Mellon University.

2 Convex Sets

2.1 Affine and Convex Sets

Definition (Affine). A set $C \subseteq \mathbb{R}^n$ is said to be *affine* if for all $x, y \in C$, we have that $\theta x + (1-\theta)y \in C$ for all $\theta \in \mathbb{R}$.

Geometrically, affine sets are sets in which the line formed by any two points in the set is entirely contained in the set as well.

Definition (Affine Combination). An *affine combination* of $\{x_i\}_{i=1}^k \subseteq \mathbb{R}^n$ is a linear combination

$$\sum_{i=1}^{k} \theta_i x_i$$

where $\theta_i \in \mathbb{R}$ for all $i \in [k]$ and $\sum_{i=1}^k \theta_i = 1$.

Proposition. Let $C \subseteq \mathbb{R}^n$ be an affine set. Then, any affine combination of $\{x_i\}_{i=1}^k \subseteq C$ is contained in C.

Proof. By induction on k.

Definition (Affine Hull). The *affine hull* of a set $C \subseteq \mathbb{R}^n$ is the set

$$\text{aff}(C) = \{\theta x + (1-\theta)y : x, y \in C, \theta \in \mathbb{R}\}.$$

Exercise. Prove that the affine hull of $C \subseteq \mathbb{R}^n$ is the smallest affine set containing C.

Affine sets are certainly unbounded, as lines are unbounded. We can consider bounded sets by considering only line segments rather than entire lines. Thus lies the idea behind convexity.

Definition (Convex). A set $C \subseteq \mathbb{R}^n$ is said to be *convex* if for all $x, y \in C$, we have that $\theta x + (1-\theta)y \in C$ for all $\theta \in [0, 1]$.

Geometrically, convex sets are sets in which the line segment formed by any two points in the set is entirely contained in the set as well. Note that we only consider line segments by restricting our coefficients to $[0, 1]$.

We can similarly extend the idea of affine combinations and affine hulls to convexity.

Definition (Convex Combination). A *convex combination* of $\{x_i\}_{i=1}^k \subseteq \mathbb{R}^n$ is a linear combination

$$\sum_{i=1}^{k} \theta_i x_i$$

where $\theta_i \in [0,1]$ for all $i \in [k]$ and $\sum_{i=1}^{k} \theta_i = 1$.

Proposition. Let $C \subseteq \mathbb{R}^n$ be a convex set. Then, any convex combination of $\{x_i\}_{i=1}^{k} \subseteq C$ is contained in C.

Proof. By induction on k.

Definition (Convex Hull). The *convex hull* of a set $C \subseteq \mathbb{R}^n$ is the set

$$\text{conv}(C) = \{\theta x + (1-\theta)y : x, y \in C, \theta \in [0,1]\}.$$

Exercise. Prove that the convex hull of $C \subseteq \mathbb{R}^n$ is the smallest convex set containing C.

2.2 Examples

2.2.1 Cones

Definition (Cone). A set C is a *cone* if for all $x \in C$, $\lambda x \in C$ for all $\lambda \geq 0$.

The traditional image of a "cone" is itself a cone, if extended to infinity.

Not all cones are convex: the union of two different lines is a cone, but not convex.

2.2.2 Hyperplanes and Halfspaces

Definition (Hyperplane). Let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. The set

$$\{x \in \mathbb{R}^n : a^T x = b\}$$

is a *hyperplane*.

Geometrically, hyperplanes are lines in $\mathbb{R}^2$ and planes in $\mathbb{R}^3$.

Proposition. All hyperplanes are affine.

Proof. Consider the hyperplane

$$S = \{x \in \mathbb{R}^n : a^T x = b\}.$$

Consider $x, y \in S$ and $\theta \in \mathbb{R}$. Then,

$$a^T\left(\theta x + (1-\theta)y\right) = \theta(a^T x) + (1-\theta)(a^T y) = \theta b + (1-\theta)b = b.$$

Hence, $\theta x + (1-\theta)y \in S$ and so S is affine.

Definition (Halfspace). Let $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. The set

$$\{x \in \mathbb{R}^n : a^T x \leq b\}$$

is a *halfspace*.

Geometrically, halfspaces are one side of a hyperplane.

Proposition. All halfspaces are convex.

2.2.3 Balls and Ellipsoids

Definition (Ball). A ball is a set

$$B(x_c, r) = \{x \in \mathbb{R}^n : ||x - x_c|| \leq r\},$$

where $r > 0$.

Proposition. All balls are convex.

Proof. Consider some ball $B(x_c, r)$. Take $x, y \in B(x_c, r)$ and $\theta \in [0, 1]$. Let $z = \theta x + (1-\theta)y$. We wish to show that $z \in B(x_c, r)$. Observe that by triangle inequality,

$$||z - x_c|| = ||\theta x + (1-\theta)y - x_c|| \leq ||\theta x - \theta x_c|| + ||(1-\theta)y - (1-\theta x_c)|| \leq \theta r + (1-\theta)r = r,$$

hence $z \in B(x_c, r)$.

Definition (Ellipsoid). An ellipsoid is a set

$$\{x \in \mathbb{R}^n : (x - x_c)^T P^{-1}(x - x_c) \leq 1\}$$

where $x_c \in \mathbb{R}^n$ and $P \in S^n_{++}$.

Proposition. All ellipsoids are convex.

Proof. Consider some ellipsoid

$$C = \{x \in \mathbb{R}^n : (x - x_c)^T P^{-1}(x - x_c) \leq 1\}.$$

Take $x, y \in C$, and $\theta \in [0,1]$. Let $z = \theta x + (1-\theta)y$. We wish to show that $z \in C$. Observe that

$$
\begin{aligned}
&(z - x_c)^T P^{-1}(z - x_c) \\
&= (\theta x + (1-\theta)y - x_c)^T P^{-1}(\theta x + (1-\theta)y - x_c) \\
&= (\theta x + (1-\theta)y - \theta x_c - (1-\theta)x_c)^T P^{-1}(\theta x + (1-\theta)y - \theta x_c - (1-\theta)x_c) \\
&= (\theta(x - x_c) + (1-\theta)(y - x_c))^T P^{-1}(\theta(x - x_c) + (1-\theta)(y - x_c)) \\
&= \theta^2 (x - x_c)^T P^{-1}(x - x_c) + \theta(1-\theta)((x - x_c)^T P^{-1}(y - x_c) + (y - x_c)^T P^{-1}(x - x_c)) \\
&+ (1-\theta)^2 (y - x_c)^T P^{-1}(y - x_c) \\
&= \theta^2 (x - x_c)^T P^{-1}(x - x_c) + 2\theta(1-\theta)(x - x_c)^T P^{-1}(y - x_c) + (1-\theta)^2 (y - x_c)^T P^{-1}(y - x_c) \\
&\leq \theta^2 (x - x_c)^T P^{-1}(x - x_c) + 2\theta(1-\theta)\sqrt{(x - x_c)^T P^{-1}(x - x_c)}\sqrt{(y - x_c)^T P^{-1}(y - x_c)} \\
&+ (1-\theta)^2 (y - x_c)^T P^{-1}(y - x_c) \\
&= \left(\theta\sqrt{(x - x_c)^T P^{-1}(x - x_c)} + (1-\theta)\sqrt{(y - x_c)^T P^{-1}(y - x_c)}\right)^2 \\
&\leq (\theta + (1-\theta))^2 \\
&= 1.
\end{aligned}
$$

Hence, $z \in C$, and so all ellipsoids are convex.

2.2.4 Polyhedra

Definition (Polyhedra). A polyhedra is a set

$$P = \{x \in \mathbb{R}^n : \big(\forall i \in [m], a_i^T x \leq b_i\big) \wedge \big(\forall i \in [p], c_i^T x = d_i\big)\}.$$

Polyhedra are typically presented with the notation

$$P = \{x \in \mathbb{R}^n : Ax \preceq b, Cx = d\}$$

where $A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, C \in \mathbb{R}^{p \times n}, d \in \mathbb{R}^p$.

Geometrically, polyhedra are the intersection of a finite number of hyperplanes and halfspaces.

2.3 Operations that Preserve Convexity

2.3.1 Intersection

Proposition. Intersection preserves convexity.

Proof. Let I be an index set such that for all i, $C_i \subseteq \mathbb{R}^n$ is a convex set. Let $C = \bigcap_{i \in I} C_i$. We claim that C is convex.

Consider $x, y \in C$, and $\theta \in [0, 1]$. Set $z = \theta x + (1 - \theta)y$. It suffices to show that $z \in C$.

By definition of C, $x, y \in C_i$ for all $i \in I$. By convexity of C_i, we have that $z \in C_i$ for all $i \in I$. Hence, $z \in C$, as desired.

Exercise. The positive semidefinite cone, S_+^n, is convex.

Solution. Observe that

$$S_+^n = \bigcap_{z \in \mathbb{R}^n} \{X \in \mathbb{R}^{n \times n} : z^T X z \geq 0\}.$$

Every set on the right hand side is convex. The intersection of convex sets is convex, hence the positive semidefinite cone is convex.

2.3.2 Affine Functions

Definition (Affine Function). A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *affine* if it is of the form

$$f(x) = Ax + b.$$

Proposition. The image of a convex set over an affine function is convex.

Proof. Let $C \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}^m$ an affine function via $f(x) = Ax + b$. We wish to show that $f[C]$ is convex.

Consider $x, y \in f[C]$, and $\theta \in [0, 1]$. Define $z = \theta x + (1 - \theta)y$. It suffices to show that $z \in f[C]$.

As $x, y \in f[C]$, there exists $s, t \in C$ such that $f(s) = x$ and $f(t) = y$.

Let $u = \theta s + (1 - \theta)t$. As C is convex, we have that $u \in C$.

Thus,

$$\begin{aligned} z &= \theta x + (1-\theta)y \\ &= \theta f(s) + (1-\theta) f(t) \\ &= \theta As + \theta b + (1-\theta)At + (1-\theta)b \\ &= A(\theta s + (1-\theta)t) + b \\ &= Au + b \\ &= f(u) \in f[C], \end{aligned}$$

as desired.

The above proposition implies that scaling, translation, and projection preserve convexity.

Proposition. The pre-image of a convex set over an affine function is convex.

Proof. Let $C \subseteq \mathbb{R}^m$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}^m$ an affine function via $f(x) = Ax + b$. We wish to show that $f^{-1}[C]$ is convex.

Consider $x, y \in f^{-1}[C]$, and $\theta \in [0, 1]$. It suffices to show that $z = \theta x + (1 - \theta)y \in f^{-1}[C]$, which in turn we must show that $f(z) \in C$.

Observe that

$$\begin{aligned} f(z) &= f\left(\theta x + (1-\theta)y\right) \\ &= \theta Ax + \theta b + (1-\theta)Ay + (1-\theta)b \\ &= \theta f(x) + (1-\theta) f(y) \in C \end{aligned}$$

by convexity of C.

2.3.3 Perspective Function

Definition (Perspective Function). The perspective function $P : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}^n$ is defined via

$$P(s, t) = \frac{s}{t}.$$

Proposition. The image of a convex set over the perspective function is convex.

Proof. Let $C \subseteq \mathbb{R}^n \times \mathbb{R}_{++}$ be a convex set. We wish to show that $P[C]$ is convex.

Consider $x, y \in P[C]$. We wish to show that for any $\theta \in [0,1]$, $\theta x + (1-\theta)y \in P[C]$. The proof is difficult if done directly, hence we will take a slightly different approach.

Fix $\theta \in [0,1]$. Since $x, y \in P[C]$, there exists $(a,s),(b,t) \in C$ such that $P(a,c) = x$ and $P(b,t) = y$.

By convexity of C, we have that

$$(\theta a + (1-\theta)b, \theta s + (1-\theta)t) = \theta(a,s) + (1-\theta)(b,t) \in C.$$

Hence,

$$\frac{\theta s P(x)}{\theta s + (1-\theta)t} + \frac{(1-\theta)tP(y)}{\theta s + (1-\theta)t} = \frac{\theta a + (1-\theta)b}{\theta s + (1-\theta)t} = P(\theta a + (1-\theta)b, \theta s + (1-\theta)t) \in P[C]$$

Let

$$\mu = \frac{\theta s}{\theta s + (1-\theta)t}.$$

Through substitution, we have that

$$\mu P(x) + (1-\mu)P(y) \in P[C].$$

As we vary $\theta \in [0,1]$, μ varies from 0 to 1, implying that $P[C]$ is convex.

Proposition. The pre-image of a convex set over the perspective function is convex.

Proof. Let $C \subseteq \mathbb{R}^n$ be a convex set. We wish to show that $P^{-1}[C]$ is convex.

Consider $(x,s),(y,t) \in P^{-1}[C]$, and $\theta \in [0,1]$. It suffices to show that $z = \theta(x,s) + (1-\theta)(y,t) \in P^{-1}[C]$. Thus, we wish to show that $P(z) \in C$.

Observe that

$$\begin{aligned} P(z) &= P\left(\theta(x,s) + (1-\theta)(y,t)\right) \\ &= P\left(\theta x + (1-\theta)y, \theta s + (1-\theta)t\right) \\ &= \frac{\theta x + (1-\theta)y}{\theta s + (1-\theta)t} \\ &= \frac{\theta s}{\theta s + (1-\theta)t} P(x,s) + \frac{(1-\theta)t}{\theta s + (1-\theta)t} P(y,t). \end{aligned}$$

Then, note that $P(x,s), P(y,t) \in C$, and

$$\frac{\theta s}{\theta s + (1-\theta)t}, \frac{(1-\theta)t}{\theta s + (1-\theta)t} \in [0,1].$$

Hence, by convexity of C,

$$P(z) = \frac{\theta s}{\theta s + (1-\theta)t} P(x,s) + \frac{(1-\theta)t}{\theta s + (1-\theta)t} P(y,t) \in C,$$

implying that $z \in P^{-1}[C]$.

2.4 Separating Hyperplanes

Theorem (Separating Hyperplane). Two non-empty and disjoint convex sets can be separated by a hyperplane.

Formally, if there are non-empty and disjoint convex sets $C, D \subseteq \mathbb{R}^n$, there exists $a \in \mathbb{R}^n$, $b \in \mathbb{R}^n$ such that

$$\forall x \in C, a^T x \leq b \quad \forall x \in D, a^T x \geq b.$$

Geometrically, there exists a hyperplane between C and D.

Proof. We only prove the theorem for the special case in which C, D are closed and bounded, hence compact.

Note that $C \times D$ is then compact. We may define the distance function $D : C \times D \to \mathbb{R}$ via $D(x,y) = ||x - y||$. By compactness, there exists $u, v \in C, D$ such that $D(u,v) = ||u - v||$ is minimized.

Define $a = v - u$ and $b = (v-u)^T(v+u)/2$. We claim that the hyperplane $\{a^T x = b\}$ separates C, D.

Assume for the sake of contradiction not. Then, without loss of generality, there exists $x \in D$ such that $a^T x < b$. Implying that

$$\begin{aligned}
& a^T x < b \\
\iff & (v-u)^T x - \frac{(v-u)^T(v+u)}{2} < 0 \\
\iff & (v-u)^T \left(x - \frac{v+u}{2} \right) < 0 \\
\iff & (v-u)^T \left(x + \frac{-v-u}{2} \right) < 0 \\
\iff & (v-u)^T \left(x + \frac{v-u}{2} - v \right) < 0 \\
\iff & (v-u)^T (x-v) + ||v-u||/2 < 0.
\end{aligned}$$

Clearly, $||v-u|| > 0$, meaning that $(v-u)^T(x-v) < 0$.

Intuitively, we can move v towards the direction $x - v$ and minimize the distance from u.

Formally, we can take the derivative

$$\frac{d}{dt} ||(v + t(x-v)) - u||^2 \Big|_{t=0} = 2(v-u)^T(x-v) < 0$$

per the above.

Thus, for some small $t > 0$, we have that

$$||(v + t(x-v)) - u|| < ||v-u||.$$

By convexity of D, $v + t(x-v) = (1-t)v + tx \in D$, hence the above is a contradiction by definition of u, v.

Definition. We say that two convex sets $C, D \subseteq \mathbb{R}^n$ are *strictly separated* if there exists $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ such that

$$\forall x \in C, a^T x < b \quad \forall x \in D, a^T x > b.$$

Example. Let $C \subseteq \mathbb{R}^n$ be a closed convex set and $x_0 \in \mathbb{R}^n$ a point not in C. Then, C and $\{x_0\}$ are strictly separated.

To see why, note that as C is closed, $\mathbb{R}^n \setminus C$ is open. Hence, we can find $r > 0$ such that $B(x_0, r) \cap C = \emptyset$. Clearly, $B(x_0, r)$ is convex. By the separating hyperplane theorem, we may find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ such that $a^T x \leq b$ for all $x \in C$, and $a^t x \geq b$ for all $x \in B(x_0, r)$. In particular, the last statement means that for any $u \in \mathbb{R}^n$ where $||u|| \leq r$,

$$a^T x_0 + a^T u = a^T (x_0 + u) \geq b.$$

The left hand side is minimized when $u = -\frac{a}{r||a||}$, hence

$$a^T x_0 - r \geq b \implies a^T x_0 \geq b + r > b.$$

Thus, the hyperplane strictly separates C and $\{x_0\}$.

We can use this result to show the following.

Proposition. Let $C \subseteq \mathbb{R}^n$ be a closed convex set, and $\mathcal{H}$ be the set of all halfspaces that contain C entirely. Then,

$$C = \bigcap \mathcal{H}.$$

Proof.

$C \subseteq \bigcap \mathcal{H}$

Trivial by definition of $\mathcal{H}$.

$\bigcap \mathcal{H} \subseteq C$

Take $x \in \bigcap \mathcal{H}$. Assume for the sake of contradiction that $x \notin C$. Then, we may find a strictly separating hyperplane between $\{x\}$ and C. Implying that $x \notin \bigcap \mathcal{H}$, a contradiction.

2.5 Supporting Hyperplanes

Definition. For a convex set $C \subseteq \mathbb{R}^n$, we say that a hyperplane

$$\{x \in \mathbb{R}^n : a^T x = a^T x_0\}$$

is a *supporting hyperplane* if $x_0 \in \partial C$ and $a^T x \leq a^T x_0$ for all $x \in C$. Geometrically, the hyperplane is tangent to a point on the boundary of C, and its halfspace contains the entirety of C.

3 Convex Functions

3.1 Basic Properties and Definitions

Definition. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if for all $x, y \in f$, $\theta \in [0, 1]$, we have that

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y).$$

If the inequality is strict, i.e.

$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta)f(y),$$

for $x \neq y$, then the function is said to be *strictly convex*.

Intuitively, convex functions are those in which the epigraph of the function (the area above the function) is a convex set.

Remark. A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if it is convex when restricted to any line in its domain, i.e. for all $v \in \mathbb{R}^n$,

$$g(t) = f(x + tv)$$

is convex.

Theorem (First Order Characterization). A function $f : \mathbb{R}^n \to \mathbb{R}$ in C^1 is convex if and only if

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

for all $x, y \in \mathbb{R}^n$.

Proof.

Case: $n = 1$

First assume that f is convex. Then for any $x, y \in \mathbb{R}$, we have that

$$
\begin{aligned}
& f(x+\theta(y-x)) \leq (1-\theta)f(x) + \theta f(y) \\
\implies & f(x+\theta(y-x)) \leq f(x) - \theta f(x) + \theta f(y) \\
\implies & f(x) + \frac{f(x+\theta(y-x)) - f(x)}{\theta} \leq f(y) \\
\implies & \lim_{\theta \to 0} f(x) + \frac{f(x+\theta(y-x)) - f(x)}{\theta} \leq f(y) \\
\implies & f(x) + \lim_{\theta \to 0} \frac{f(x+\theta(y-x)) - f(x)}{\theta} \leq f(y) \\
\implies & f(x) + (y-x) \lim_{\theta \to 0} \frac{f(x+\theta(y-x)) - f(x)}{\theta(y-x)} \leq f(y) \\
\implies & f(x) + f'(x)(y-x) \leq f(y).
\end{aligned}
$$

as desired.

Now instead assume that for all $x, y \in \mathbb{R}$,

$$f(x) + f'(x)(y-x) \leq f(y).$$

We wish to show that f is convex. Fix $x, y \in \mathbb{R}$ and $\theta \in [0,1]$. Let $z = \theta x + (1-\theta)y$. Then,

$$
\begin{aligned}
& f(z) + f'(z)(x-z) \leq f(x) \\
\implies & f(z) + (1-\theta)f'(z)(x-y) \leq f(x).
\end{aligned}
$$

Similarly, we can see that

$$f(z) + \theta f'(z)(y-x) \leq f(y).$$

Combining these,

$$f(z) = \theta f(z) + (1-\theta)f(z) \leq \theta f(x) + (1-\theta)f(y),$$

as desired.

Case: $n > 1$

We can study the one-dimensional function which varies x in the direction of $y - x$, i.e. $g(t) = f(x + t(y - x))$. The result then follows by the $n = 1$ case.

Remark. The above inequality is strict if and only if the function is strictly convex.

Corollary. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function, and $x \in \mathbb{R}^n$ a point such that $\nabla f(x) = 0$. Then, x is a global minimizer of f.

Proof. Consider some $y \in \mathbb{R}^n$. Per the first order characterization of convex functions,

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

Theorem (Second Order Characterization). A function $f : \mathbf{dom}(f) \to \mathbb{R}$ in C^2 is convex if and only if

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

for all $x \in \mathbf{dom}(f)$.

Proof.

First assume that f is convex. Assume for the sake of contradiction that there exists some $x \in \mathbf{dom}(f)$ such that $\nabla^2 f(x)$ is not positive semi-definite. By definition, there exists some eigenvector $v \in \mathbb{R}^n$ such that $\nabla^2 f(x)v = \lambda v$ where $\lambda < 0$.

Define $g(t) = f(x + tv)$. Note that

$$g''(t) = v^T \nabla^2 f(x + tv) v.$$

Hence,

$$g''(0) = v^T \nabla^2 f(x) v = v^T \lambda v = \lambda ||v||^2 < 0.$$

By definition of the second derivative, for some small $\epsilon > 0$, $g'(c) < g'(0)$ for all $c \in (0, \epsilon)$.

Then,

$$g(\epsilon) = g(0) + (g(\epsilon) - g(0)) = g(0) + g'(c)\epsilon < g(0) + g'(0)\epsilon,$$

by the mean value theorem. Hence, a contradiction since g inherits convexity from f and the above violates the first order characterization of convexity. Thus, $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbf{dom}(f)$.

Now assume that $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbf{dom}(f)$. We claim that f is convex.

Fix $x, y \in \mathbf{dom}(f)$. Define

$$g(t) = f(x + t(y - x)).$$

Then,

$$f(y) = g(1) = g(0)+(g(1)-g(0)) = g(0)+g'(c) \geq g(0)+g'(0) = f(x)+\nabla f(x)^T(y-x)$$

by mean value theorem. The inequality follows from the fact that g'' is always non-negative, implying that g' is non-decreasing.

Since x, y are arbitrary, we have that f is convex by the first order characterization of convexity.

Remark. The above inequality is strict if and only if the function is strictly convex.

3.2 Operations that Preserve Convexity

3.2.1 Nonnegative Weighted Sum

Proposition. Let $\{f_i\}_{i=1}^n$ be a sequence of convex functions. Then,

$$\sum_{i=1}^{n} \omega_i f_i, \quad \omega_i \geq 0$$

is convex.

Proof. Fix $x, y \in \mathbb{R}^n$ and $\theta \in [0, 1]$. Then,

$$\begin{aligned}\left(\sum_{i=1}^{n}\omega_i f_i\right)(\theta x+(1-\theta y)) &= \sum_{i=1}^{n}\omega_i f_i(\theta x+(1-\theta)y)\\ &\leq \sum_{i=1}^{n}\omega_i\left(\theta f_i(x)+(1-\theta)f_i(y)\right)\\ &= \theta\left(\sum_{i=1}^{n}\omega_i f_i(x)\right)+(1-\theta)\left(\sum_{i=1}^{n}\omega_i f_i(y)\right)\\ &= \theta\left(\sum_{i=1}^{n}\omega_i f_i)\right)(x)+(1-\theta)\left(\sum_{i=1}^{n}\omega_i f_i\right)(y)\end{aligned}$$

as desired.

Remark. The above proposition generalizes to infinite sums (if they converge) as well as integrals. Specifically, if a function $f(x,y)$ is convex in x for all $y \in A$, then

$$g(x)=\int_A \omega(y)f(x,y)dy, \quad \omega(y)\geq 0$$

is convex.

3.2.2 Affine Composition

Proposition. Let $f:\mathbb{R}^n\to\mathbb{R}$, $A\in\mathbb{R}^{n\times m}$, and $b\in\mathbb{R}^n$. Then,

$$g(x)=f\left(Ax+b\right)$$

is convex if f is convex.

Proof. Fix $x,y\in\mathbb{R}^n$ and $\theta\in[0,1]$. Then,

$$\begin{aligned}g\left(\theta x+(1-\theta)y\right) &= f\left(A\left(\theta x+(1-\theta)y\right)+b\right)\\ &= f\left(\theta(Ax+b)+(1-\theta)(Ay+b)\right)\\ &\leq \theta f(Ax+b)+(1-\theta)f(Ay+b)\\ &= \theta g(x)+(1-\theta)g(y).\end{aligned}$$

3.2.3 Maximum and Supremum

Proposition. Let $\{f_i\}_{i=1}^n$ be a sequence of convex functions. Then,

$$g(x) = \max_i f_i(x)$$

is convex.

Proof. Fix $x, y \in \mathbb{R}^n$ and $\theta \in [0,1]$. Then,

$$\begin{aligned} g\left(\theta x + (1-\theta)y\right) &= \max_i f_i\left(\theta x + (1-\theta)y\right) \\ &\leq \max_i \left(\theta f_i(x) + (1-\theta) f_i(y)\right) \\ &\leq \theta \max_i f_i(x) + (1-\theta) \max_i f_i(y) \\ &= \theta g(x) + (1-\theta) g(y). \end{aligned}$$

Remark. The above proposition generalizes to the supremum. Specifically, if $f(x,y)$ is convex in x for all y, then

$$g(x) = \sup_y f(x,y)$$

is convex.

3.2.4 Representation as Supremum of Affine Functions

Proposition. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then,

$$f(x) = \sup\{g(x) : g \text{ affine}, g(y) \leq f(y) \quad \forall y \in \mathbb{R}^n\}.$$

Proof.

$f(x) \leq \sup\{g(x) : g \text{ affine}, g(y) \leq f(y) \quad \forall y \in \mathbb{R}^n\}$

Define

$$\mathbf{epi}(f) = \{(x,y) : f(x) \leq y\}.$$

We claim that $\mathbf{epi}(f)$ is convex. Fix $(x_1, y_1), (x_2, y_2) \in \mathbf{epi}(f)$, and $\theta \in [0,1]$. Then let $z = \theta(x_1, y_1) + (1-\theta)(x_2, y_2)$. So,

$$z = (\theta x_1 + (1-\theta)x_2, \theta y_1 + (1-\theta)y_2).$$

By convexity of f,

$$f(\theta x_1 + (1-\theta)x_2) \leq \theta f(x_1) + (1-\theta)f(x_2) \leq \theta y_1 + (1-\theta)y_2$$

implying that $z \in \mathbf{epi}(f)$, as desired. Hence, $\mathbf{epi}(f)$ is convex.

Now fix some $x \in \mathbb{R}^n$. We shall show that indeed,

$$f(x) \leq \sup\{g(x) : g \text{ affine}, g(y) \leq f(y) \quad \forall y \in \mathbb{R}^n\}.$$

Observe that $(x, f(x)) \in \partial\mathbf{epi}(f(x))$, hence we may find $a \in \mathbb{R}^n$, $b \in \mathbb{R}$ such that

$$\begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} x - z \\ f(x) - t \end{bmatrix} \leq 0$$

for all $(z, t) \in \mathbf{epi}(f)$.

Then,

$$\begin{aligned} &a^T(x-z) + b(f(x) - z) \leq 0 \\ \implies &a^T(x-z) + b(f(x) - f(z) - s) \leq 0 \end{aligned}$$

for all s. Implying that $b > 0$.

We can then write the above as

$$g(z) = \frac{a^T}{b}(x-z) + f(x) \leq f(z)$$

for all $z \in \mathbb{R}^n$.

Hence, g is an affine function that underestimates f over all z, and achieves $g(x) = f(x)$. We thus have the result.

$f(x) \geq \sup\{g(x) : g \text{ affine}, g(y) \leq f(y) \quad \forall y \in \mathbb{R}^n\}$

Follows by definition.

3.2.5 Perspective of a Function

Definition. Consider a function $f : \mathbb{R}^n \to \mathbb{R}$. We define the perspective of f as $g_f : \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}$ via

$$g_f(x, t) = tf(x/t).$$

Proposition. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then g_f is convex.

Proof. It suffices to show that the epigraph of g_f is convex. Note that the epigraph of g_f is the preimage of the epigraph of f over the perspective function, hence is convex.

3.3 Convex Conjugates

Definition. Let $f : \mathbb{R}^n \to \mathbb{R}$. The *conjugate* of f, $f^\star : \mathbb{R}^n \to \mathbb{R}$ is defined via

$$f^\star(y) = \sup_{x \in \mathbf{dom}(f)} \{y^T x - f(x)\}.$$

Geometrically, the conjugate of a function f is the greatest distance between f and the hyperplane $y^T x$.

Proposition. For any function $f : \mathbb{R}^n \to \mathbb{R}$, the conjugate $f^\star$ is always convex.

Proof. Observe that $y^T x - f(x)$ is convex in y, hence $f^\star(y) = \sup_{x \in \mathbf{dom}(f)} \{y^T x - f(x)\}$ is convex in y.

3.3.1 Basic Properties

Proposition (Fenchel's Inequality). Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function. Then,

$$f(x) + f^\star(y) \geq x^T y$$

for all x, y.

Proof. Note that

$$f^\star(y) \geq x^T y - f(x)$$

for all x, y. The result is then immediate.

4 Convex Optimization Problems

4.1 Optimization Problems

Definition. An *optimization problem* is of the form

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p. \end{aligned}$$

Its domain is the intersection of domains for each function, i.e.

$$D = \bigcap_{i=0}^{m} \mathbf{dom}(f_i) \cap \bigcap_{i=1}^{p} \mathbf{dom}(h_i).$$

Definition. The *feasible set* of an optimization problem is the set

$$\Omega = \{x \in D : f_i(x) \leq 0, h_i(x) = 0\}.$$

We say that an optimization problem is feasible if its feasible set is not empty.

Definition. We define the optimal value of an optimization problem as the value

$$p^\star = \inf_{x \in \Omega} f_0(x).$$

If $\Omega = \emptyset$, then $p^\star = \infty$.

Definition. If there exists a sequence $\{x_i\}_{i=1}^{\infty} \subseteq \Omega$ such that $f_0(x_k) \to -\infty$ as $k \to \infty$, we say that the optimization problem is *unbounded below* and $p^\star = -\infty$.

Definition. We say that $x \in \Omega$ is ϵ - *suboptimal* if $f_0(x) \leq p^\star + \epsilon$.

Definition. We say that $x \in \Omega$ is *locally optimal* if there exists some $R > 0$ such that $f_0(z) \geq f_0(x)$ for all $z \in B(x, R)$.

Definition. If $f_i(x) = 0$ for some $i \in [m]$ and $x \in \Omega$, we say that constraint i is *active* at x.

Definition. We call an optimization problem of the form

$$\begin{aligned} \min_x \quad & 0 \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p. \end{aligned}$$

a *feasibility problem.*

Remark. Note that maximization problem can be formulated as optimization problems by taking $-f_0$.

4.1.1 Slack Variables

Slack variables allow us to express inequalities as equalities.

In particular, note that

$$f_i(x) \leq 0 \iff f_i(x) + \xi = 0$$

for some $\xi \geq 0$ (in particular, $\xi = -f_i(x)$). Here, ξ is a *slack variable.*

More generally, we can reformulate the optimization problem

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p. \end{aligned}$$

as

$$\begin{aligned} \min_{x,\xi} \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) + \xi_i = 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, p \\ & \xi_i \geq 0, \quad i = 1, \dots, m. \end{aligned}$$

4.2 Convex Problems

Definition. A *convex problem* is an optimization problem of the form

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & a_i^T x = b, \quad i = 1, \ldots, p. \end{aligned}$$

where f_i is convex for all i.

Remark. The feasible set of a convex optimization problem

$$\Omega = \bigcap_{i=1}^{m} \{x : f_i(x) \leq 0\} \cap \mathbf{dom}(f_0) \cap \{x : Ax = b\}$$

is convex as it is the intersection of convex sets.

Proposition. Any local solution to a convex optimization problem is a global solution.

Proof. Let $x \in \Omega$ be a local solution to the typical convex optimization problem. Then, there exists $R > 0$ such that x is optimal in the R ball around it.

Assume for the sake of contradiction that x is not locally optimal. Then, there exists some $y \in \Omega$ such that $f_0(y) < f_0(x)$.

We can find some $\theta \in [0, 1]$ such that $x + \theta(y - x) \in B(x, R)$. Then,

$$f_0\left(x + \theta(y - x)\right) \leq (1 - \theta) f_0(x) + \theta f_0(y) < f_0(x),$$

a contradiction.

The above proposition provides some intuition as to why convex optimization problems are particularly nice to work with.

Proposition. Let $f_0 \in C^1(\Omega)$ be a convex function. Then, $x \in \Omega$ is optimal if and only if

$$\nabla f_0(x)^T (y - x) \geq 0$$

for all $y \in \Omega$.

Proof. First assume that $x \in \Omega$ is optimal. Assume for the sake of contradiction that $\nabla f_0(x)^T(y-x) < 0$ for some y. Define

$$g(t) = f(x + t(y-x)).$$

Observe then that

$$g'(0) = \nabla f_0(x)^T(y-x) < 0.$$

Thus, for some $\epsilon > 0$, for all $c \in (0, \epsilon)$, we have that $g(c) < g(0)$, implying that

$$f(x + c(y-x)) < f(x),$$

a contradiction.

Now assume that

$$\nabla f_0(x)^T(y-x) \geq 0$$

holds for all x. We claim that x is optimal. Observe that

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

as desired.

Corollary. Consider some convex optimization problem where $f_0 \in C^1(\Omega)$. If Ω is open and $x \in \Omega$ is the optimal point, then

$$\nabla f(x)^T(y-x) = 0$$

for all $y \in \Omega$.

Proof.

As Ω is open, we can find small enough $\theta > 0$ such that $y = x - \theta \nabla f_0(x) \in \Omega$. Then,

$$-\theta ||\nabla f_0(x)||^2 = \nabla f_0(x)^T((x - \theta \nabla f_0(x)) - x) \geq 0$$

which is only true if the gradient is zero.

Proposition: Consider the convex optimization problem

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ \text{s.t.} \quad & Ax = b \end{aligned}$$

where $f_0 \in C^1(\Omega)$. Then, a point $x^\star$ is an optimal point if and only if

$$\nabla f_0(x^\star) + A^T v = 0$$

for some v, and

$$Ax^\star = b.$$

Proof. We first prove the forwards direction. Assume that $x^\star$ is optimal. Then, $Ax^\star = b$ trivially. Furthermore, we know that for all $y \in \Omega$,

$$\nabla f_0(x^\star)^T (y - x^\star) \geq 0.$$

As $y \in \Omega$, we know that $Ay = b$ as well. Thus, we must have that $y = x + v$ where $v \in \mathcal{N}(A)$. We may then rewrite the above as

$$\nabla f_0(x^\star)^T v \geq 0$$

for all $v \in \mathcal{N}(A)$.

Thus, $\nabla f_0(x^\star)$ is orthogonal to $\mathcal{N}(A)$, hence $\nabla f_0(x^\star) \in \mathcal{R}(A^T)$. Similarly, its negative is in the range of A^T. Meaning that there must exist some v such that

$$\nabla f_0(x^\star) + A^T v = 0.$$

We now prove the backwards direction.

If $Ax^\star = b$, then clearly $x^\star$ is feasible.

If

$$\nabla f_0(x^\star) + A^T v = 0,$$

then $\nabla f_0(x^\star) \in \mathcal{R}(A^T)$. Hence, it is orthogonal to $\mathcal{N}(A)$, and so

$$\nabla f_0(x^\star)^T (y - x^\star) = 0$$

for all $y \in \Omega$.

Thus, $x^\star$ is optimal.

4.3 Linear Problems

Definition. A *linear problem* (LP) is an optimization problem of the form

$$\begin{array}{ll} \min\limits_x & c^T x \\ \text{s.t.} & Ax = b \\ & x \succeq 0. \end{array}$$

One may introduce inequalities in the constraints via slack variables.

4.4 Quadratic Problems

Definition. A *quadratic problem* (QP) is an optimization problem of the form

$$\begin{array}{ll} \min\limits_x & \frac{1}{2} x^T P x + q^T x + r \\ \text{s.t.} & Gx \preceq h \\ & Ax = b, \end{array}$$

where $P \in S_+^n$.

Definition. A *quadratically constrained quadratic problem* (QCQP) is an optimization problem of the form

$$\begin{aligned} \min_x \quad & \frac{1}{2}x^T P_0 x + q_0^T x + r_0 \\ \text{s.t.} \quad & \frac{1}{2}x^T P_i x + q_i^T x + r_i \leq 0 \\ & Ax = b, \end{aligned}$$

where $P_i \in S_+^n$.

Definition. A *second-order cone program* (SOCP) is an optimization problem of the form

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & ||A_i x + b_i||_2 \leq c_i^T x + d_i \\ & Fx = g. \end{aligned}$$

5 Duality

5.1 Lagrange Dual Function

Definition. The *Lagrangian* of a (not necessarily convex) optimization problem

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p. \end{aligned}$$

is a function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ defined via

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \lambda^T f(x) + \nu^T h(x).$$

Definition: The *Lagrange dual* of an optimization problem is the function $g : \mathbb{R}^m \times \mathbb{R}^p$ defined via

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

where $\mathcal{L}$ is the Lagrangian of the optimization problem and D is the domain of the optimization problem (not necessarily feasible).

Proposition. The Lagrange dual of any optimization problem is concave.

Proof. Observe that the Lagrangian is affine in λ, ν. Concavity is preserved under point-wise infimum.

Proposition (Weak Duality). For some optimization problem, let $p^\star$ be its optimal value and $g(\lambda, \nu)$ be its Lagrange dual. Then, for any $\lambda \geq 0$ and any ν,

$$g(\lambda, \nu) \leq p^\star.$$

Proof. Fix $\lambda \succeq 0$ and ν. Consider some feasible x. Then,

$$\begin{aligned} g(\lambda, \nu) &\leq \mathcal{L}(x, \lambda\nu) \\ &= f_0(x) + \lambda^T f(x) + \nu^T h(x) \\ &= f_0(x) + \lambda^T f(x) \\ &\leq f_0(x) \\ &\leq p^\star \end{aligned}$$

as desired.

Definition. If $(\lambda, \nu) \in \mathbf{dom}(g)$ and $\lambda \succeq 0$, then we say that (λ, ν) is *dual feasible*.

5.2 Lagrange Dual Problem

In light of weak duality, we can think of finding the best lower bound on the optimal value using the Lagrange dual function. Thus is the motivation for the Lagrange dual problem, defined below. Note that the Lagrange dual problem is particularly nice due to the fact that the Lagrange dual function is concave, established previously.

Definition. Let $g(\lambda, \nu)$ be the Lagrange dual for an optimization problem. We define the *Lagrange dual problem* for the optimization problem as

$$\begin{aligned} \max_{\lambda,\nu} \quad & g(\lambda, \nu) \\ \text{s.t.} \quad & \lambda \succeq 0 \end{aligned}$$

We call this problem the *dual*, and the original problem the *primal*.

Remark. Let $p^\star$ be the optimal value for the primal problem, and $d^\star$ be the optimal value to the dual problem. Then, by weak duality,

$$d^\star \leq p^\star.$$

There may be more implicit constraints, particularly when g is unbounded below (recall that it is an infimum).

Definition. The *optimal duality* gap of a problem is the value

$$p^\star - d^\star \geq 0.$$

Definition. We say that *strong duality* holds if

$$p^\star = d^\star.$$

5.3 Geometric Intuition

We now build some geometric intuition regarding duality.

Fix some optimization problem and define the set

$$G = \{(f(x), h(x), f_0(x)) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} | x \in D\}.$$

G essentially expresses all value combinations of the constraints and objective. We now interpret many prior results, along with some new results, using the geometry of G.

5.3.1 Optimal Value

It is easy to see that

$$p^\star = \inf\{t : (u, v, t) \in G, u \leq 0, v = 0\},$$

i.e. we restrict the set we consider to only feasible points.

5.3.2 Lagrange Dual Function

We can also see that

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \lambda^T f(x) + \nu^T h(x) = (\lambda, \nu, 1)^T (f(x), h(x), f_0(x))$$

and

$$g(\lambda, \nu) = \inf_{x \in D} \mathcal{L}(x, \lambda, \nu) = \inf_{x \in D} (\lambda, \nu, 1)^T (f(x), h(x), f_0(x)) = \inf\{(\lambda, \nu, 1)^T (u, v, t) | (u, v, t) \in G\}.$$

Thus, for any $(u, v, t) \in G$, we have that

$$(\lambda, \nu, 1)^T (u, v, t) \geq g(\lambda, \nu).$$

If the infimum in g is attained, we can think of $(\lambda, \nu, 1), g(\lambda, \nu)$ as a supporting hyperplane to G.

5.3.3 Weak Duality

Say that $\lambda \geq 0$. Then,

$$\begin{aligned} p^\star &= \inf\{t : (u, v, t) \in G, u \leq 0, v = 0\} \\ &\geq \inf\{(\lambda, \nu, 1)^T (u, v, t) : (u, v, t) \in G, u \leq 0, v = 0\} \\ &\geq \inf\{(\lambda, \nu, 1)^T (u, v, t) : (u, v, t) \in G\} \\ &\geq g(\lambda, \nu). \end{aligned}$$

Thus, for any λ, ν where $\lambda \geq 0$, we have that $p^\star \geq g(\lambda, \nu)$. As $d^\star$ is the maximum value of $g(\lambda, \nu)$ over all λ, ν where $\lambda \geq 0$, we thus have that

$$p^\star \geq d^\star,$$

proving weak duality.

5.3.4 Epigraph Variation

Define

$$A = \{(u, v, t) | \exists x \in D, f_i(x) \leq u_i, h_i(x) = v_i, f_0(x) \leq t\}.$$

A can be thought of as sort of an epigraph of G, with the exception that we enforce equality on the equality constraints h.

Once again, the optimal value can be expressed as

$$p^\star = \inf\{t : (0, 0, t) \in A\}.$$

For $\lambda \geq 0$, note that

$$g(\lambda, \nu) = \inf\{(\lambda, \nu, 1)^T (u, v, t) | (u, v, t) \in G\} = \inf\{(\lambda, \nu, 1)^T (u, v, t) | (u, v, t) \in A\},$$

as G is a subset of A, but points in A do not decrease the value of $(\lambda, \nu, 1)^T (u, v, t)$.

Once again, we may say that if the infimum is attained, then $(\lambda, \nu, 1), g(\lambda, \nu)$ is a supporting hyperplane to A since for all $x \in A$, we have

$$(\lambda, \nu, 1)^T x \geq g(\lambda, \nu).$$

Note that $(0, 0, p^\star)$ is in the boundary of A, hence

$$p^\star = (\lambda, \nu, 1)^T (0, 0, p^\star) \geq g(\lambda, \nu)$$

once again gives us weak duality.

5.3.5 Slater's Condition

Proposition (Slater's Condition). If there exists an "interior" to the inequality constraints of a convex optimization problem, i.e.

$$\exists x \;\; f_i(x) < 0 \quad \forall i = 1, \ldots, m$$

and x is feasible, then strong duality holds.

Proof. We will use the geometric interpretation of duality to prove Slater's condition.

First note that strong duality holds if and only if

$$p^\star = (\lambda, \nu, 1)^T (0, 0, p^\star) = g(\lambda, \nu)$$

for some λ, ν. In other words, there exists a supporting hyperplane $(\lambda, \nu, 1), g(\lambda, \nu)$ to A (defined above) with $\lambda \geq 0$ tangent to $(0, 0, p^\star)$.

The idea behind the proof is that we will separate A from the set

$$B = \{(0, 0, s) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} | s < p^\star\}$$

with a hyperplane that proves strong duality.

In doing so, we will make the following simplifying assumptions: $\mathbf{int}(D) \neq \emptyset$, $\mathbf{rank}(A) = p$, and $p^\star > -\infty$ (otherwise $d^\star = -\infty = p^\star$ by weak duality).

Note that as we are only considering a convex optimization problem, A is convex as it is the Cartesian product of convex sets.

Furthermore, see that

$$A \cap B = \emptyset$$

since $p^\star$ is optimal.

As A, B are convex and disjoint, we can separate them. More precisely, there exists $(\lambda, \nu, \mu) \neq 0, \alpha$ such that

$$(\lambda, \nu, \mu)^T (u, v, t) \geq \alpha \quad (u, v, t) \in A$$

and

$$(\lambda, \nu, \mu)^T (0, 0, s) \leq \alpha \quad (0, 0, s) \in B.$$

From the first inequality, note that $\lambda \succeq 0, u \geq 0$, otherwise we could scale the left-hand side to negative infinity.

We can rewrite the last inequality as

$$\mu s \leq \alpha \quad s < p^\star,$$

meaning that

$$\mu p^\star \leq \alpha.$$

Thus,

$$\begin{aligned}\sum_{i=1}^{m} \lambda_i u_i + \sum_{i=1}^{p} \nu_i v_i + \mu t &= \lambda^T u + \nu^T v + \mu^T t \\ &= (\lambda, \nu, \mu)^T (u, v, t) \\ &\geq \alpha \\ &\geq \mu p^\star.\end{aligned}$$

For now, assume that $\mu > 0$. We will address the $\mu = 0$ case later.

Dividing both sides by μ, we have that

$$\mathcal{L}(x, \lambda/\mu, \nu/\mu) \geq p^\star$$

for all x (recall that (u, v, t) was an arbitrary element of A).

We can then minimize x over the left-hand side to recover

$$g(\lambda/\mu, \nu/\mu) \geq p^\star.$$

Weak duality, however, grants us

$$g(\lambda/\mu, \nu/\mu) = p^\star.$$

Hence, when $\mu > 0$, we have strong duality.

We now consider the $\mu = 0$ case. Then,

$$\begin{aligned}\sum_{i=1}^{m}\lambda_i u_i + \sum_{i=1}^{p}\nu_i v_i &= \sum_{i=1}^{m}\lambda_i u_i + \sum_{i=1}^{p}\nu_i v_i + \mu t \\ &= (\lambda,\nu,\mu)^T(u,v,t) \\ &\geq \alpha \\ &\geq \mu p^\star \\ &= 0\end{aligned}$$

As (u, v, t) is an arbitrary element of A, we have that for all $x \in D$,

$$\sum_{i=1}^{m}\lambda_i f_i(x) + \nu^T(Ax - b) \geq 0.$$

Then let x be a Slater point. Plugging this x into the above inequality, we have that

$$\sum_{i=1}^{m}\lambda_i f_i(x) \geq 0,$$

but $f_i(x) < 0$ for all i. Hence, $\lambda_i \leq 0$. But, $\lambda \succeq 0$, thus $\lambda = 0$.

Returning to the original inequality, we have that for all $x \in D$,

$$\nu^T(Ax - b) \geq 0.$$

Note that $\nu \neq 0$ as $(\lambda, \nu, \mu) \neq 0$ but $\lambda, \mu = 0$.

Let x once again be the Slater point. Then, we have that $\nu \neq 0$ but

$$\nu^T(Ax - b) = 0.$$

As x is in the interior, there must exist some other point $y \in D$ such that

$$\nu^T(Ay - b) < 0$$

unless $\nu^T A = 0$. But, we stated that $\mathbf{rank}(A) = p$, hence we have a contradiction. Thus, $\mu \neq 0$ and strong duality holds by the other case.

5.4 Optimality Conditions

5.4.1 Certificate of Suboptimality

The dual function provides us with a method to "certify" the suboptimality of a solution. In particular, say that we are given a solution x to some optimization problem and wish to provide a guarantee on how suboptimal it is. We can use a dual solution (λ, ν) as our certificate. By weak duality, we have that

$$g(\lambda, \nu) \leq p^\star \leq f_0(x).$$

Hence,

$$f_0(x) - p^\star \leq f_0(x) - g(\lambda, \nu).$$

Thus, our certificate tells us that the suboptimality is at most

$$f_0(x) - g(\lambda, \nu),$$

which is typically called the duality gap.

5.4.2 Complementary Slackness

Proposition (Complementary Slackness). Consider some optimization problem in which strong duality holds. Let $x^\star$ be primal optimal and $(\lambda^\star, \nu^\star)$ be dual optimal. Then for all $i = 1, \ldots, m$,

$$\lambda_i^\star f_i(x^\star) = 0.$$

Proof. Observe that

$$\begin{aligned} f_0(x^\star) &= g(\lambda^\star, \nu^\star) \\ &= \inf_{x \in D} \{f_0(x) + \lambda^{\star T} f(x) + \nu^{\star T} h(x)\} \\ &\leq f_0(x^\star) + \lambda^{\star T} f(x^\star) + \nu^{\star T} h(x^\star) \\ &= f_0(x^\star) + \lambda^{\star T} f(x^\star) \\ &\leq f_0(x^\star), \end{aligned}$$

meaning that

$$f_0(x^\star) + \lambda^{\star T} f(x^\star) = f_0(x^\star),$$

i.e.

$$\sum_{i=1}^{m} \lambda_i^\star f_i(x^\star) = 0.$$

As $\lambda^\star \succeq 0$ and $f_i(x^\star) \leq 0$, we immediately recover complementary slackness.

One can also see from the equalities that $x^\star$ is the minimizer to $\mathcal{L}(x, \lambda^\star, \nu^\star)$.

5.4.3 KKT Conditions

Definition. Consider an optimization problem in which $f_0, f_1, \ldots, f_m, h_1, h_2, \ldots, h_p$ are differentiable and strong duality holds. The *KKT conditions* for primal solution x and dual solutions (λ, ν) refer to the following conditions:

- x is primal feasible.
- (λ, ν) is dual feasible.
- $\lambda_i f_i(x) = 0$ for all $i = 1, \ldots, m$.
- $\nabla_x \mathcal{L}(x, \lambda, \nu) = 0$.

Proposition. Consider an optimization problem with the above conditions. Furthermore, let $x^\star$ be an optimal primal solution and $(\lambda^\star, \nu^\star)$ be an optimal dual solution. Then, these optimal points satisfy the KKT conditions.

Proof. It is clear that $x^\star$ is primal feasible and $(\lambda^\star, \nu^\star)$ is dual feasible. Complementary slackness holds from earlier. Furthermore, see that

$$\inf_{x \in D} \mathcal{L}(x, \lambda^\star, \nu^\star) = \mathcal{L}(x^\star, \lambda^\star, \nu^\star)$$

meaning that

$$\nabla_x \mathcal{L}(x^\star, \lambda^\star, \nu^\star) = 0,$$

as desired.

Proposition. Consider a *convex* optimization problem with differentiable functions. Furthermore, let $x, (\lambda, \nu)$ be points that satisfy the KKT conditions. Then, x is primal optimal, (λ, ν) is dual optimal, and strong duality holds.

Proof. Clearly, x is primal feasible and (λ, ν) is dual feasible by the KKT conditions. We now show optimality.

As we are considering a convex optimization problem, see that $\mathcal{L}(x, \lambda, \nu)$ is convex in x. Hence, if the gradient vanishes for any x, that x must be a global minimizer. Implying that

$$g(\lambda, \nu) = \mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) = f_0(x)$$

by invoking complementary slackness.

As the point has the dual equal to the primal, there is zero optimality gap, implying that x is primal optimal and (λ, ν) is dual feasible. Furthermore, strong duality holds.

6 Unconstrained Minimization

We now discuss methods to solve *unconstrained* minimization problems, i.e. problems of the form

$$\min_{x \in \mathbb{R}^n} f(x)$$

where f is convex and twice differentiable.

As we are in the unconstrained setting, a point $x^\star$ is optimal if and only $\nabla f(x^\star) = 0$. There are typically no analytical solutions. Hence, the general idea of these algorithms is to iteratively solve for such a point, i.e. find a sequence $\{x^{(k)}\}_{k=1}^{n}$ such that $\nabla f(x^{(k)}) \to 0$ and consequently $f(x^{(k)}) \to p^\star$.

6.1 Strong Convexity

Definition. We say that a function f is *strongly convex* on S if there exists $m > 0$ such that

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

which means that $\nabla^2 f(x) - mI \succeq 0$. We also say that f is m strongly convex.

Proposition. Let f be an m strongly convex function on S. Then, for all $x, y \in S$,

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

i.e. we have stronger guarantees on the first-order characterization of convexity.

Proof. By Taylor's and mean value theorem, there exists some z on the line segment $[x, y]$ such that

$$f(y) = f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x).$$

By m strong convexity,

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

as desired.

Proposition. Let f be an m strongly convex function on S and $p^\star$ the minimum value of f. Then for any $x \in S$,

$$p^\star \geq f(x) - \frac{m}{2}||\nabla f(x)||_2^2.$$

In words, one can bound the suboptimality of a point using its gradient.

Proof. We know that for all $y \in S$,

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

We will find $\tilde{y}$ that minimizes the right-hand side, and we have that it serves as a lower-bound for any y on the left-hand side.

The right-hand side is a convex function in y, hence we take the gradient and solve for 0:

$$\nabla f(x) + m(\tilde{y} - x) = 0 \implies \tilde{y} = x - \frac{1}{m}\nabla f(x).$$

Plugging this in,

$$f(y) \geq f(x) - \frac{1}{m}||\nabla f(x)||_2^2 + \frac{1}{2m}||\nabla f(x)||_2^2 = f(x) - \frac{1}{2m}||\nabla f(x)||_2^2.$$

This holds for any $y \in S$, hence we set $y = x^\star$ and see that

$$p^\star \geq f(x) - \frac{1}{2m}||\nabla f(x)||_2^2$$

as desired.

Corollary. Let f be an m strongly convex function on S and $p^\star$ the minimum value of f. Then for any $x \in S$, if

$$||\nabla f(x)||_2^2 \leq (2m\epsilon)^{1/2},$$

we have that

$$f(x) - p^\star \leq \epsilon.$$

We can also bound a point's distance from the minimizer using the gradient.

Proposition. Let f be an m strongly convex function on S and $x^\star$ the minimizer of f. Then for any $x \in S$,

$$||x^\star - x|| \leq \frac{2}{m}||\nabla f(x)||_2.$$

Proof. By the first-order characterization of m strong convexity,

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

via Cauchy Schwarz.

As $p^\star \leq f(x)$,

$$0 \geq -||\nabla f(x)||_2||x^\star - x||_2 + \frac{m}{2}||x^\star - x||_2^2$$

hence

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

meaning that

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

as desired.

6.1.1 Smoothness

Strong convexity imposes a lower bound on the Hessian of a function. We can similarly impose an upper bound.

Definition. We say that a function f is M *smooth* on S if

$$\nabla^2 f(x) \leq MI$$

for all $x \in S$.

Proposition. Let f be an M smooth function. Then for all $x, y \in S$,

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{M}{2}||y - x||_2^2.$$

Proof. Once again by Taylor's and mean value theorem, we have that

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T\nabla^2 f(z)(y - x)$$

for some z on the line segment $[x, y]$.

By M smoothness, we then have that

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} ||y - x||_2^2.$$

Proposition. Let f be an M smooth function with optimal value $p^\star$. Then for any $x \in S$,

$$p^\star \leq f(x) - \frac{1}{2M} ||\nabla f(x)||_2^2.$$

Proof. We will employ a similar strategy as in the convexity case, with some changes. We know that for all $y \in S$,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} ||y - x||_2^2.$$

We first find $\tilde{y}$ which *minimizes* the right-hand side. From before, we found that

$$\tilde{y} = x - \frac{1}{m} \nabla f(x).$$

Plugging this in,

$$p^\star \leq f(\tilde{y}) \leq f(x) - \frac{1}{2M} ||\nabla f(x)||_2^2.$$

6.2 Conditioning

Definition (Condition Number). Consider an unconstrained optimization problem with an objective that is m strongly convex and M smooth. We call $K = M/m$ the *condition number* of the problem.

Definition (Width). We define the *width* of a set C in direction q with unit-norm as

$$W(C, q) = \sup_{z \in C} q^T z - \inf_{z \in C} q^T z.$$

Definition. We define the *maximum* width of a set C as

$$W_{max} = \sup_{q, ||q||_2 = 1} W(C, q).$$

We define the *minimum* width of a set C as

$$W_{min} = \inf_{q, ||q||_2 = 1} W(C, q).$$

Definition. The *condition number* of a set C is

$$\mathbf{cond}(C) = \frac{W_{max}^2}{W_{min}^2}.$$

Definition. The α *sublevel set* of f is the set

$$C_\alpha = \{x : f(x) \leq \alpha\}.$$

Proposition. Consider a function that is m strongly convex and M smooth. Then, for any α,

$$\mathbf{cond}(C_\alpha) \leq K = M/m.$$

Proof. Observe that by the first-order characterizations,

$$p^\star + \frac{m}{2}||y - x^\star||_2^2 \leq f(y) \leq p^\star + \frac{M}{2}||y - x^\star||_2^2.$$

Hence, defining

$$B_{inner} = \{y | ||y - x^\star||_2 \leq (2(\alpha - p^\star)/M)^{1/2}\}$$

and

$$B_{outer} = \{y | ||y - x^\star||_2 \leq (2(\alpha - p^\star)/m)^{1/2}\}$$

we have that

$$B_{inner} \subseteq C_\alpha \subseteq B_{outer}.$$

Dividing the squared radii of the balls, we have an upper-bound on the condition number of C_α:

$$\mathbf{cond}(C_\alpha) \leq \frac{M}{m}$$

as desired.

6.3 Descent Methods

Descent methods are algorithms that solve unconstrained minimization problems by iteratively computing a direction to perturb the current solution, and the scale of said direction. Formally, on iteration $k+1$, they update the current solution via

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$$

where $t^{(k)} > 0$. $t^{(k)}$ and $\Delta x^{(k)}$ are chosen such that $f(x^{(k+1)}) < f(x^{(k)})$, i.e. we gradually aproach the optimal solution.

Note that by the first-order characterization of convexity, we require that

$$\nabla f(x^{(k)})^T \Delta x^{(k)} < 0,$$

otherwise there is no hope of finding a more optimal solution.

How the direction $\Delta x^{(k)}$ is computed depends on the specific algorithm. Several methods exist to compute $t^{(k)}$, some of which are covered below.

6.3.1 Exact Line Search

Simply set $t^{(k)}$ such that the objective is minimized:

$$t^{(k)} = \arg\min_{s>0} f(x^{(k)} + s\Delta x^{(k)}).$$

While it is true that we must solve another optimization problem, this problem is one-dimensional and in practice is very easy to solve.

6.3.2 Backtracking

A simplified backtracking algorithm to compute t would be to initially set $t^{(k)}$ to 1, then while $f(x^{(k)} + t^{(k)}\Delta x^{(k)}) > f(x^{(k)})$, halve $t^{(k)}$.

6.4 Gradient Descent

Gradient descent provides one natural option to choose the descent direction: the negative of the gradient. Until the stopping criterion is satisfied (e.g. the current gradient is small enough), we update the current solution via

$$x^{(k+1)} = x^{(k)} - t^{(k)}\nabla f(x^{(k)})$$

where $t^{(k)}$ is computed by either exact line search or backtracking.

Proposition. Consider some unconstrained minimization problem over f, where f is m strongly convex and M smooth. If we were to perform gradient descent with exact line search beginning at $x^{(0)}$, we would reach an ϵ optimal solution at step N where

$$N \leq \frac{\log((f(x^0) - p^\star)/\epsilon)}{\log(1/c)},$$

note that $c = 1 - m/M$.

Proof. By M smoothness and the first-order characterization, we have that

$$f(x - t\nabla f(x)) \leq f(x) - t||\nabla f(x)||_2^2 + \frac{Mt^2}{2}||\nabla f(x)||_2^2.$$

As we use exact line search, we can improve our bound by finding t that minimizes the right-hand side, which is convex. We take the gradient with respect to t and set it to 0:

$$0 = -||\nabla f(x)||_2^2 + Mt||\nabla f(x)||_2^2 \implies t = \frac{1}{M}.$$

Substituting this in,

$$f(x - t\nabla f(x)) \leq f(x) - \frac{||\nabla f(x)||_2^2}{M} + \frac{||\nabla f(x)||_2^2}{2M} \leq f(x) - \frac{||\nabla f(x)||_2^2}{2M}.$$

Hence,

$$f(x - t\nabla f(x)) - p^\star \leq (f(x) - p^\star) - \frac{||\nabla f(x)||_2^2}{2M},$$

i.e. we always improve our optimality gap by

$$\frac{1}{2M}||\nabla f(x)||_2^2.$$

Recall that with m strong convexity, the gradient provides us with a bound on our suboptimality:

$$f(x) - p^\star \leq \frac{1}{2m}||\nabla f(x)||_2^2.$$

Hence, we can restate our bound as

$$f(x - t\nabla f(x)) - p^\star \leq (f(x) - p^\star) - \frac{m}{M}(f(x) - p^\star) = c(f(x) - p^\star).$$

By induction,

$$f(x^{(k)}) - p^\star \leq c^k(f(x^{(0)}) - p^\star).$$

To have that $f(x^{(k)} - p^\star) \leq \epsilon$, it is sufficient to have

$$c^k \leq \frac{\epsilon}{f(x^{(0)}) - p^\star}$$

meaning that it is sufficient for

$$k \leq \frac{\log\left(\frac{\epsilon}{f(x^{(0)}) - p^\star}\right)}{\log(c)} = \frac{\log\left((f(x^{(0)}) - p^\star)/\epsilon\right)}{\log(1/c)}$$

as desired.