Convex Optimization: Theory (CS5580)

Instructor: Saketh

Contents

Lecture 1

Introduction & Definition of MP

1. To highilight the importance and non-triviality involved in converting an informal description of an optimization problem in English to a formal one using the Language of Mathematics, the following two examples were presented:

   (a) Given m points $\mathcal{D} = \{x_1, \ldots, x_m\} \subset \mathbb{R}^n$, find the sphere with smallest volume that encloses all the points in $\mathcal{D}$:

   $$\begin{aligned} \min_{c\in\mathbb{R}^n, r\in\mathbb{R}^{++}} \quad & \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} r^n \\ \text{s.t.} \quad & \|x_i - c\| \leq r \ \forall\ i = 1, \ldots, m. \end{aligned} \tag{1.1}$$

   (b) Given m points $\mathcal{D} = \{x_1, \ldots, x_m\} \subset \mathbb{R}^n$, find the ellipsoid[1] with smallest volume that encloses all the points in $\mathcal{D}$:

   $$\begin{aligned} \min_{c\in\mathbb{R}^n, \Sigma\succ 0} \quad & \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} det(\Sigma)^{\frac{-1}{2}} \\ \text{s.t.} \quad & (x_i - c)^\top \Sigma (x_i - c) \leq 1 \ \forall\ i = 1, \ldots, m. \end{aligned} \tag{1.2}$$

2. Motivated by above examples, we defined a Mathematical Program (MP): A symbol of the following form is defined as a MP:

   $$\begin{aligned} \min_{x\in\mathcal{X}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \ \forall\ i = 1, \ldots, m. \end{aligned} \tag{1.3}$$

[1] We recalled the positive-definite (pd) matrices (if M is pd, we denote it by $M \succ 0$). Refer http://en.wikipedia.org/wiki/Positive-definite_matrix. It is helpful if one is familiar with all results pertaining to pd matrices, especially the one about its eigen value decomposition: $M \succ 0 \Leftrightarrow M = L\Lambda L^\top$, where L is an orthonormal matrix and Λ is a diagonal matrix with positive entries. The entries in the diagonal matrix are called eigen-values and columns in the orthonormal matrix can be taken as the corresponding eigen-vectors.

3. It was easy to verify that (1.1), and (1.2), are both MPs.

4. We defined x as the variable, $\mathcal{X}$ as the domain, $f : \mathcal{X} \mapsto \mathbb{R}_{ext}$ as[2] the objective (function), the inequalities $g_i(x) \leq 0$ as the constraints, the functions $g_i : \mathcal{X} \mapsto \mathbb{R}_{ext}$ are called as the constraint functions, the set $\mathcal{F} \equiv \{x \in \mathcal{X} \mid g_i(x) \leq 0 \; \forall \; i = 1, \ldots, m\}$ as the feasibility set, each member of the feasibility set is called as a feasible solution/point, for the MP (1.3).

5. The value of the MP (1.3) is defined as the $\inf(\{f(x) \mid x \in \mathcal{F}\})$, with the understanding that the value is defined as $-\infty$ if the set of feasible function values, $\{f(x) \mid x \in \mathcal{F}\}$, is not bounded below, and is defined as ∞ if the feasibility set is empty. This value is also sometimes called as the optimal value.

6. We clarified that in this course we will be interested in MPs where the variables live in Euclidean Spaces. This goes by the name continuous optimization[3].

[2]$\mathbb{R}_{ext}$ denotes the extended real numbers i.e., $\mathbb{R} \cup \{-\infty, \infty\}$. The reason for including $\pm\infty$ will be clear when discussing cascaded MPs etc.

[3]The MPs where variables are members of discrete sets, are studied in typical CS algorithms, and popular as discrete optimization. Also we restrict ourselves to finite dimensional spaces.

Lecture 2

Parameterized & Cascaded MPs

1. If an MP's value is $-\infty$, i.e., the set of values of the objective function over the feasibility set is not bounded below, then the MP is said to be unbounded. If an MP's value is ∞, i.e., the feasibility set is empty, then the MP is said to be infeasible.

2. We looked at many examples of MPs, and watched out for all the above definitions. In particular, we realized that the form (1.3) is "universal". Specifically, for the sake of convenience, constraints may sometimes be pushed into the domain's ($\mathcal{X}$) definition[1].

3. Associated with every MP of form (1.3), there is a related program of the form:

$$\begin{aligned} \arg\min_{x \in \mathcal{X}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \ \forall \ i = 1, \ldots, m. \end{aligned} \tag{2.1}$$

   (a) All the definitions of variable, domain, objective, constraints, feasibility set remain the same.

   (b) The value of (2.1) is defined as the set $\{x \in \mathcal{F} \mid f(x) = V\}$, where V is the value of the original MP (1.3). i.e., it is the set of all feasible points where the infimum value is attained by the objective. This set is also known as the Solution set of the original MP (1.3) and its members are called as the solutions of (1.3). The phrase "Solution" is also sometimes qualified as "optimal/optimal solution" or as "global optimal/optimum solution" or sometimes simplified as the "optimal/optimum".

[1] or vice-versa if that is still well-defined.

(c) An MP is said to be solvable iff its solution set is non-empty. Further, it is uniquely solvable iff its solution set is a single-ton.

4. We thought of examples of MPs with empty, single-ton,..., countably infinite, uncountably infinite, solution sets.

5. We defined an alternative but equivalent form for an MP:

$$\begin{aligned} \max_{x\in\mathcal{X}} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0 \; \forall \; i = 1, \dots, m. \end{aligned} \tag{2.2}$$

   (a) All the definitions of variable, domain, objective, constraints, feasibility set remain the same.

   (b) The value of (2.2) is defined as the $\sup\left(\{f(x) \mid x \in \mathcal{F}\}\right)$, with the understanding that the value is defined as ∞ if the set of feasible function values, $\{f(x) \mid x \in \mathcal{F}\}$, is not bounded above, and is defined as $-\infty$ if the feasibility set is empty.

   (c) Since $\inf(S) = -\sup(-S)$, both forms[2] are equivalent.

6. The notion of value naturally defines a total order on the set of all MPs:

   (a) We say MP1=MP2 iff the value of MP1 is equal to the value of MP2.

   (b) We say MP1>MP2 iff the value of MP1 is greater than the value of MP2.

   (c) By $MP1 = 2$, we mean the value of MP1 is 2.

7. This helped us to write down functions of MPs: $h(MP)$ is nothing but h evaluated at the value of the MP. We commented that if h is a monotonically-non-decreasing continuous, then:

$$h\left(\begin{array}{cc} \min_{x\in\mathcal{X}} & f(x) \\ \text{s.t.} & g_i(x) \le 0 \; \forall \; i \end{array}\right) = \begin{array}{cc} \min_{x\in\mathcal{X}} & h(f(x)) \\ \text{s.t.} & g_i(x) \le 0 \; \forall \; i \end{array}$$

8. We then noted that many a time, the objective and constraint functions may involve other parameters, which we call as the parameters of the MP. Such a parameterized version of a MP is very useful for studying a collection of MPs that only differ in parameter values. The general form for a parameterized MP is:

$$\begin{aligned} \min_{x\in\mathcal{X}} \quad & f(x, y) \\ \text{s.t.} \quad & g_i(x, y) \le 0 \; \forall \; i = 1, \dots, m, \end{aligned} \tag{2.3}$$

[2] $-S$ is the set with members as the negatives of those in S.

where the $y \in \mathcal{Y}$ are the parameters. We noted examples of such parameterized MPs.

(a) It is easy to see that the value of the MP (2.3) depends on the parameters y. Hence we denote the value of the parametrized MP by a function of y:

$$h(y) \equiv \min_{x \in \mathcal{X}} \quad f(x, y) \quad \text{s.t.} \quad g_i(x, y) \leq 0 \ \forall\ i = 1, \ldots, m,$$

(b) Now one can deal with h as any other function. For e.g., one may want to intergrate, or differentiate, or compose it with other functions. The most interesting operation on h is an optimization (over y). We first looked at:

$$\min_{y \in \mathcal{Y}} h(y) \equiv \min_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \quad f(x, y) \quad \text{s.t.} \quad g_i(x, y) \leq 0 \ \forall\ i = 1, \ldots, m, \tag{2.4}$$

(c) It is an easy exercise to show that:

Theorem 2.0.1.

$$\begin{array}{cccc} \min_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} & f(x,y) & = & \min_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} & f(x,y) \\ \text{s.t.} & g_i(x,y) \leq 0\ \forall\ i & & \text{s.t.} & g_i(x,y) \leq 0\ \forall\ i. \\ & & = & \min_{z\in\mathcal{Z}} & f(z) \\ & & & \text{s.t.} & g_i(z) \leq 0\ \forall\ i, \end{array}$$

where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}, z = (x, y)$.

In this sense, such cascaded MPs behave just like multiple integrals (order of integration does'nt matter).

(d) In the next lecture we will look at $\max_{y \in \mathcal{Y}} h(y)$.

Lecture 3

Review of Vector Spaces

1. We continued the discussion on cascaded MPs, by considering:

$$\begin{aligned} \max_{y\in\mathcal{Y}} h(y) \equiv \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \quad & f(x,y) \\ \text{s.t.} \quad & g_i(x,y) \le 0 \ \forall\ i = 1,\dots,m, \end{aligned} \tag{3.1}$$

2. Using various examples, we illustrated that:

$$\begin{array}{cc} \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} & f(x,y) \\ \text{s.t.} & g_i(x,y) \le 0 \ \forall\ i \end{array} \overset{??}{=} \begin{array}{cc} \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} & f(x,y) \\ \text{s.t.} & g_i(x,y) \le 0 \ \forall\ i, \end{array}$$

where $\overset{??}{=}$ can be $\ll$ or $<$ or $=$.

3. We then proved[1] the following theorem:

Theorem 3.0.1.

$$\begin{array}{cc} \max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} & f(x,y) \\ \text{s.t.} & g_i(x,y) \le 0 \ \forall\ i \end{array} \le \begin{array}{cc} \min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} & f(x,y) \\ \text{s.t.} & g_i(x,y) \le 0 \ \forall\ i, \end{array}$$

A simple proof appears here: `https://en.wikipedia.org/wiki/Maxmin_inequality`.

4. Note that in the cascaded/recursive MPs (2.4,3.1), essentially the objective function (of the outer MP) was defined using an MP (the inner MP). Needless to say, one can also explore the possibility where the constraint function(s) is(are) defined using an MP. For e.g., let $g_i(x) \equiv \min_{y\in\mathcal{Y}} h_i(y)$, then:

$$\begin{array}{cc} \min_{x\in\mathcal{X}} & f(x) \\ \text{s.t.} & g_i(x) \le 0 \ \forall\ i \end{array} \equiv \begin{array}{cc} \min_{x\in\mathcal{X}} & f(x) \\ \text{s.t.} & \min_{y\in\mathcal{Y}} h_i(y) \le 0 \ \forall\ i, \end{array}$$

[1] All proofs will either appear in hand-written notes or appropriate references will be explicitly cited in these notes.

5. We then argued that understanding the special structure implied by the objective function, feasibility set, and more fundamentally, the underlying space in which the variable lives, is important to understand the nature of the associated MP. Hence we began with a review of vector spaces (since in continuous optimization, the variables are assumed to live in Euclidean spaces)[2]:

   (a) Vector space is formally defined in page 9 of Sheldon Axler [1997].

   (b) We gave examples of vector spaces — Euclidean, those with matrices, functions etc.

   (c) We identified linear combination as an important operation (V is closed under linear combinations by the axioms).

   (d) We asked if every set of vectors V, has a subset of vectors, say B, such that linear span of B, $LIN(B) \equiv \{\sum_{i=1}^{m} \lambda_i v_i \mid \lambda_i \in \mathbb{R},\ v_i \in B\ \forall\ i = 1, \ldots, m, m \in \mathbb{N}\}$, i.e., the set of all vectors which can be expressed as linear combinations of those in B, is equal to V? Obviously such sets exist (for example take $B = V$ itself). Such sets are called as the spanning sets of V.

   (e) A vector space is finite-dimensional if there exists a spanning set of finite size. In this course we will restrict ourselves to finite-dimensional ones.

   (f) We said that it will be great if i) the spanning set is small (smallest). (Then the proposed representation will be highly compact) ii) the proposed representation is one-to-one.

   (g) We argued[3] that answer to both goals is the same: a Basis, which is a linearly independent, spanning set. A linearly independent set is a set of vectors whose non-trivial (not all zero) linear combination can never give a trivial vector (zero vector).

   (h) Theorem 2.6 in Sheldon Axler [1997] says that cardinality of a linearly independent set is always lesser than that of a spanning set. From this it easily follows that cardinality of any basis of a vector space is the same. Hence basis is indeed the smallest spanning set. The common size of any basis is called the dimensionality of the vector space.

   (i) Hence a basis is like a pair of goggles, through which the **vector space** looks "simple". We noted that every finite dimensional vector space has a basis and is essentially equivalent to a Euclidean vector space of same dimensionality. Also, a basis gives an inner/constitutional/compositional/primal

[2] Go through pages 1–13 in [Sheldon Axler, 1997]. Also go through related exercises.
[3] Refer pages 21-36 in [Sheldon Axler, 1997].

description (a description of an object with help of parts in it) of the vector space it spans.

(j) For the vector space examples, we noted a basis in each case, and computed the dimensionality.

(k) Given a vector space $\mathcal{V} = (V, +, \cdot)$, there exists subsets $W \subseteq V$, which themselves form a vector space: $\mathcal{W} \equiv (W, +, \cdot)$ — such a subset is called a linear set or linear variety and the resulting vector space is called a subspace[4] of the original vector space. In lectures, we may interchangeably use the terms subspace and linear set (as long as it does'nt create much confusion).

(l) We studied some examples of subspaces in various vector spaces and noted their basis.

(m) Euclidean vector spaces are interesting not only because of lin. comb., but also because notions of dot-products, distances, projections and other such interesting operations exist. In order to make abstract vector spaces interesting, we defined a new operator $< \cdot, \cdot >: V \times V \mapsto \mathbb{R}$, called the inner-product, which satisfies positive-definiteness, symmetry and linearity properties and extends the idea of a dot-product in Euclidean spaces[5]. A vector space endowed with a valid inner-product is called an inner-product space.

[4] It is important to note that the operators in the subspace are the same as that in the original vector space.

[5] Refer pg 98-101 of [Sheldon Axler, 1997] for definition and examples.

Lecture 4

Review of Inner-Product Spaces

1. We gave many examples of inner-products with Euclidean vectors, Matrices. In particular, we noted that $\langle v, w\rangle_M \equiv v^\top M w (M \succ 0)$ is the general form of inner-products in Euclidean spaces (and hence analogous in any finite dimensional space). Since M induces the entire geometry (as we will see later), it is called the kernel.

2. Innerproduct naturally induces a notion of orthogonality: $v \perp w \iff \langle v, w\rangle = 0$. We noted how notion of orthgonality changes with the kernel. In particular, we noted that in the usual matrix space, the set of symmetric and skew-symmetric matrices are orthogonal[1]. More interestingly, we noted that symmetric and skew-symmetric matrices form sub-spaces whose dimensionalties i.e., $\frac{n(n+1)}{2}, \frac{n(n-1)}{2}$, add up to n. Such (inner-product) subspaces are said to be orthogonal complements of each other.

3. We then noted the induced notion of angle: we defined angle $\angle v, w \equiv \arccos\left(\frac{\langle u,v\rangle}{\sqrt{\langle u,u\rangle}\sqrt{\langle v,v\rangle}}\right)$. We then proved the Cauchy-Schwartz inequality[2], which implied that the angle formula is well-defined.

4. We the defined induced norm $\|v\| \equiv \sqrt{\langle v, v\rangle}$. Then we showed (again using Cauchy-Schwartz inequality) that the induced norm is a valid norm[3].

5. Once the notion of norm is also in place, the notions of distance between vectors, projections onto vectors/sets, orthogonal basis, geometric figures like sphere, parallelogram, cube, conic-sections, etc. naturally follow. Also,

[1] Sets are orthogonal if each pair of members from the sets are orthogonal.

[2] See 6.6 in Sheldon Axler [1997] for a proof (different from that done in the lecture, but very insightful).

[3] See 6.2-6.12 in Sheldon Axler [1997] for definition of norm etc.

one can prove other basis geometric results like Pythgorean, Parallelogram theorem etc.

6. Also, analysis definitions like Cauchy, convergent sequence, limits naturally follow. An inner-product space that is complete (all Cauchy sequences converge) is called a Hilbert space. In this course we will be concerned with variables living in finite-dimensional Hilbert spaces.

Lecture 5

Linear Sets

1. We noted[1] that every finite-dimensional Hilbert space has a finitely sized orthogonal/orthonormal basis, i.e., a basis who members have unit-norm and every pair of members are orthogonal to each other.

2. We showed how using an orthogonal basis, one can show an equivalence between any finite-dimensional Hilbert space and Euclidean space of equal dimension.

   (a) Let $B = \{v_1, ..., v_n\}$ be an orthonormal basis of the n-dimensional Hilbert space, $\mathcal{H} = (V,\ +_V,\ \cdot_V,\ \langle\cdot,\cdot\rangle_V)$. Let $u = \sum_{i=1}^n \lambda_i v_i$ and $w = \sum_{i=1}^n \alpha_i v_i$. Since B is a basis, $v = w \iff \lambda_i = \alpha_i\ \forall\ i = 1, ..., n$. This proves that the map $v \mapsto \lambda \equiv \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix}$ is bijective.

   (b) Interestingly, $\rho_1 \cdot_V v +_V \rho_2 \cdot_V w \mapsto \rho_1 \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} + \rho_2 \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$. This shows the equivalence of the linear combinations.

   (c) More interestingly, $\langle v, w\rangle_V = \lambda^\top \alpha$. This gives the equivalence between the inner-products.

   (d) Hence all theorems/results we know in Euclidean spaces must hold in any finite-dimensional Hilbert space.

3. We talked about an operation called direct summing that will enable us to join a Hilbert space of say Euclidean vectors with that of say matrices: Given two inner-product/Hilbert spaces $\mathcal{H}_1 = (V_1,\ +_1,\ \cdot_1,\ \langle\cdot,\cdot\rangle_1)$ and

[1] Cor. 6.24 in Sheldon Axler [1997].

$\mathcal{H}_2 = (V_2,\ +_2,\ \cdot_2,\ \langle\cdot,\cdot\rangle_2)$, we dened the direct sum of those, $\mathcal{H} \equiv \mathcal{H}_1 \bigoplus \mathcal{H}_2 \equiv (V_*,\ +_*,\ \cdot_*,\ \langle\cdot,\cdot\rangle_*)$, where $V_* \equiv V_1 \times V_2$ (cartesian product), and for any $v = (v_1, v_2), w = (w_1, w_2) \in V_*$, where $v_1, w_1 \in V_1, v_2, w_2 \in V_2$, we have $v +_* w \equiv (v_1 +_1 w_1, v_2 +_2 w_2), \alpha \cdot_* v \equiv (\alpha \cdot_1 v_1, \alpha \cdot_2 v_2)$, and $\langle v, w\rangle_* \equiv \langle v_1, w_1\rangle_1 + \langle v_2, w_2\rangle_2$. It is an easy exercise to show that the direct sum is a well-defined Hilbert space. This is the natural way of stacking up arbitrary spaces to form big space. Note that with such a direct sum, the following two subspaces are orthogonal complements of each other: $S_1 = \{(v_1, 0_2) \mid v_1 \in V_1\}$ and $S_2 = \{(0_1, v_2) \mid v_2 \in V_2\}$, where $0_1, 0_2$ denote the additive identity elements $\mathcal{H}_1, \mathcal{H}_2$ respectively. More importantly, if $A = \{a_1, \ldots, a_n\}$ is an orthonormal basis of $\mathcal{H}_1$, and $B = \{b_1, \ldots, b_m\}$ is an orthonormal basis of $\mathcal{H}_2$, then $C \equiv \{(a_1, 0_2), \ldots, (a_n, 0_2), (0_1, b_1), \ldots, (0_1, b_m)\}$ is an orthonormal basis of $\mathcal{H}$. Hence $dim(\mathcal{H}) = dim(\mathcal{H}_1) + dim(\mathcal{H}_2)$, justifying the name direct sum.

4. We noted[2] norms other than the induced norms.

5. We then began looking at special (sub)sets in Hilbert spaces (all through we assume $\mathcal{V} = (V, +, \cdot, \langle\cdot,\cdot\rangle)$ is the underlying (finite-dimensional) Hilbert space).

6. We started with the familiar Linear Sets (L): sets that are closed under linear combinations[3], i.e., $L = LIN(L)$. We call this the primal definition/characterization of linear sets. Needless to say, Basis of L is[4] the minimal way of representing L using the notion of linear combinations. We say B is the primal/inner representation/description of L.

7. We realized that every linear set can also be described using the notion of orthogonality. Let L be a linear set and B be a basis of subspace induced by it. Let us define the orthogonal complement of a set S, as, $S^\perp \equiv \{v \in \mathcal{V} \mid \langle l, v\rangle = 0\ \forall\ l \in S\}$. The following statements are true:

    (a) $L^\perp$ is a linear set (follows from linearity of the inner product). Infact, $L^\perp$ is a linear set even if L is not a linear (but an arbitrary) set. Let us denote the basis of the subspace induced by $L^\perp$ as $B^\perp$. Needless to say, $B^\perp \cap B = \{0\}$.

[2] Refer https://en.wikipedia.org/wiki/Norm_(mathematics), https://en.wikipedia.org/wiki/Matrix_norm.

[3] Refer sections A.1.2, A.1.3, A.1.4, A.2.3, A.3.4 in Nemirovski [2005] for material on Linear Sets.

[4] More precisely, B is the basis of the subspace induced by L.

(b) By the rank-nullity theorem[5], it follows that $dim(L) + dim(L^{\perp}) = dim(V)$[6]. From this and orthogonality, it follows that $B^{\perp} \cup B$ is a basis for V.

(c) A key result in duality is:

Theorem 5.0.1. $L = \left(L^{\perp}\right)^{\perp}$, *whenever L is a linear set.*

Note this need not be true if L is not a linear set, in which case $L \subset \left(L^{\perp}\right)^{\perp}$.

(d) From the above it follows that $L = \left\{v \in \mathcal{V} \mid \langle l, v\rangle = 0 \; \forall \; l \in B^{\perp}\right\}$. We call this as the dual definition/characterization of a linear set. We say $B^{\perp}$ is the dual/outer representation/description of L. $B^{\perp}$ is also known as the dual basis of L. In particular, linear sets are nothing but solution sets of a system of homogeneous linear equations.

(e) $L_1 \subset L_2 \Rightarrow L_2^{\perp} \subset L_1^{\perp}$.

8. If $dim(L) \leq \left\lfloor \frac{n}{2} \right\rfloor$, then one would describe L as $LIN(B)$, else one would describe L as $\left\{v \in V \mid \langle v, b\rangle = 0 \; \forall \; b \in B^{\perp}\right\}$. Thus one would at the maximum require $\left\lfloor \frac{n}{2} \right\rfloor$ vectors to represent any Linear set!

9. We named the special linear set of dimensionality one less than the vector space as Hyperplane (through the origin). For e.g., line in 2-d, plane in 3-d etc. It was immediate that the dual definition is better suited for a hyperplane: $\mathbb{H}_w \equiv \{x \mid \langle w, x\rangle = 0\}$, where $w \neq 0$. It follows that all linear sets apart from V are either hyperplanes (through the origin) or their intersections.

---

[5]Theorem 3.4 in Sheldon Axler [1997].

[6]this justifies the name complement!

Lecture 6

Linear Sets: Calculus & Topology

1. We proved the key duality result for linear sets: $L = \left(L^{\perp}\right)^{\perp}$.

2. We discussed[1] operations that preserve linearity of sets:

   (a) Given an arbitrary collection of sets $S_\lambda,\ \lambda \in \Lambda$, where Λ is the index set[2], we define (arbitrary) intersection, $\cap_{\lambda\in\Lambda} S_\lambda \equiv \{x \mid x \in S_\lambda\ \forall\ \lambda \in \Lambda\}$. It is easy to see that (arbitrary) intersection of linear sets is linear.

   (b) Given an arbitrary collection of sets $S_\lambda,\ \lambda \in \Lambda$, where Λ is the index set[3], we define (arbitrary) union, $\cup_{\lambda\in\Lambda} S_\lambda \equiv \{x \mid \exists\lambda \in \Lambda\ \ni x \in S_\lambda\}$. It is easy to give counter examples where union of two linear sets is not linear.

   (c) Given sets $S_1, \ldots, S_n$ and reals $\lambda_1, \ldots, \lambda_n$, we define their linear combination as $\sum_{i=1}^{n} \lambda_i S_i \equiv \{\sum_{i=1}^{n} \lambda_i v_i \mid v_i \in S_i,\ \forall\ i = 1, \ldots, n\}$. It is easy to show that linear combinations of linear sets are same as a simple summation of the same sets, and are linear sets. Infact, $LIN\left(S_1 \cup S_2\right) = S_1 + S_2$.

   (d) If L is linear, then it's complement, $L^c \equiv \{v \in V \mid v \notin L\}$, will not be linear (infact L^c will not even contain 0).

   (e) Given two Linear sets L_1, L_2, their Cartesian product, $L_1 \times L_2 \equiv \{(v_1, v_2) \mid v_1 \in L_1,\ v_2 \in$ is also a linear set[4].

[1] We encourage readers to think about two different proof strategies henceforth. One based on primal definition, and the other based on dual.

[2] Index set could be finite, countably infinite or uncountable.

[3] Index set could be finite, countably infinite or uncountable.

[4] This is a sub-result used in proving that Direct sum is a valid Hilbert space.

(f) Given two sets S_1, S_2, we define their set difference as $S_1 \backslash S_2 \equiv \{v_1 \in S_1 \mid v_1 \notin S_2\}$. Again, $L_1 \backslash L_2$ will not be linear for linear L_1, L_2 (infact $L_1 \backslash L_2$ will not even contain 0).

3. We introduced some topological notions:

Closure: Given a set, S, closure[5], $Cl(S)$, is defined as the set comprised of the limits of all convergent sequences formed with elements of S.

Closed set: S is closed iff $S = Cl(S)$.

Interior Point: Given a set, S, a point $x \in S$ is said to be an interior point of S iff $B_\epsilon(x) \subseteq S$ for some $\epsilon > 0$, where $B_\epsilon(x) \equiv \{v \in V \mid \|v - x\| \leq \epsilon\}$ is the ball of radius ϵ centered at x.

Interior: The set of all interior points of S is defined as the interior, $int(S)$. A set is said to have interior iff its interior is non-empty.

Boundary: Given a set S, boundary, $\delta(S) \equiv Cl(S) \backslash int(S)$.

Bounded Set: A set S is bounded iff $B_r(0) \subseteq S$ for some finite $r > 0$.

Compact: A set S is compact iff it is closed and bounded.

4. Here are some standard results in topology:

   (a) Complementarity of open and closed sets: S is closed if and only if S^c is open.

   (b) (arbitrary) Intersection of closed sets is closed; (arbitrary) union of open sets is open.

   (c) Finite Union of closed sets is closed and finite intersection of open sets is open.

   (d) (arbitrary) intersection of bounded sets is bounded. Finite union of bounded sets is bounded.

5. Linear sets are closed[6].

6. Linear sets, except the entire set of vectors, are not open. But as we will see later, they are relatively open.

7. Linear sets, except the one containing only 0, are not bounded.

[5] B.1.6.A. in Nemirovski [2005].

[6] As all finite dimensional spaces are equivalent to Euclidean space, which is complete.

Lecture 7

Affine Sets

1. We defined affine sets as shifted linear sets: A is affine[1] iff there exists a linear set L and $a \in V$, such that $A = \{a\} + L$.

2. We defined affine combination as linear combination with the restriction that the combining coefficients sum to unity.

3. We defined affine hull:

$$AFF(S) \equiv \left\{ \sum_{i=1}^{m} \lambda_i v_i \mid \lambda_i \in \mathbb{R},\ v_i \in S\ \forall\ i = 1, \ldots, m, \sum_{i=1}^{m} \lambda_i = 1, m \in \mathbb{N} \right\},$$

   i.e., the set of all vectors which can be expressed as affine combinations of those in the set.

4. We proved that A is affine iff $A = AFF(A)$, which we took as the primal definition/characterization of Affine sets. It was easy to define notions of affinely spanning set, affine independence and affine basis (refer section A.3 in Nemirovski [2005] for all related discussions/proofs). We will call affine basis as the primal/inner representation/description.

5. We defined dimension, $dim(A) \equiv dim(L)$, which turned out to be one less that the number of elements in the affine basis.

6. We proved the dual characterization/definition: A is affine with associated linear set as L, with $B^{\perp} = \{b_1, \ldots, b_m\}$ as the basis for $L^{\perp}$, iff there exist numbers $\alpha_i \in \mathbb{R},\ i = 1, \ldots, m$, such that $A = \{v \mid \langle v, b_i \rangle = \alpha_i,\ \forall\ i = 1, \ldots, m\}$. In particular, this shows that affine sets are nothing but solution sets of

[1] Please refer sections A.3, A.4 in Nemirovski [2005] and optionally, section 1 in Rockafellar [1996] for material on Affine sets.

(non-homogeneous) linear equations. We call $(B^{\perp}, \alpha)$ as the dual/outer representation/description of A, where α is the vector with entries as α_i.

7. We call affine sets of dimensionality one less than the highest, as Hyperplane. Needless to say, the dual characterization is the most efficient: $\mathbb{H}_w \equiv \{x \mid \langle w, x \rangle = b\}$, where $w \neq 0, b \in \mathbb{R}$. It follows that all affine sets, apart from V, are either hyperplanes or their intersections.

8. We gave examples of affine sets, hyperplanes, and identified their primal and dual representations.

9. The operations that preserve affinity and the topology remains analogous to linear sets.

Lecture 8

Cones

1. We defined conic combination as linear combination with the restriction that the combining coefficients must be non-negative.

2. We defined conic hull:

$$CONIC(S) \equiv \left\{ \sum_{i=1}^{m} \lambda_i v_i \mid \lambda_i \in \mathbb{R}^+,\ v_i \in S\ \forall\ i = 1, \ldots, m, m \in \mathbb{N} \right\},$$

i.e., the set of all vectors which can be expressed as conic combinations of those in the set.

3. We say that K is a cone/conic-set iff $K = CONIC(K)$, which we took as the primal definition/characterization of Conic sets.

4. We say S is a conicly spanning set of K iff $K = CONIC(S)$. We realized examples of cones with finitely sized conicly spanning sets, which we henceforth call as Polyhedral Cones. We also saw examples like the ice-cream cone (in 3d) and the psd cone (in space of Symmetry matrices), that are NOT polyhedral cones. In each case we identified a "minimal" conicly spanning set:

   (a) For the ice-cream cone (in 3d), a minimally conicly spanning set is the unit circle at unit height.

   (b) For the psd cone, a minimally conicly spanning set is the set of all symmetric-rank-one matries i.e., matrices of the form $xx^\top,\ x \in \mathbb{R}^n$.

Lecture 9

Cones: Duality & Algebra

1. We then generalized the notion of an orthogonal complement, and defined the dual cone, S^*, of a set S: $S^* \equiv \{v \in V \mid \langle v, s\rangle \geq 0 \; s \in S\}$. It is an easy exercise to show that S^* is indeed a cone for any set S. We gave examples of dual cones, and noted that the ice-cream and psd cones are dual to themselves and hence are called as self-dual cones.

2. We proved that S^* is always a closed set.

3. We them attempted proving an important duality result:

   Theorem 9.0.1. *For a closed cone K, we have $K = (K^*)^*$.*

   While it was easy to see that $K \subseteq (K^*)^*$, we said it is not straightforward to show the converse. We noted that a separation theorem, which we will state and prove in coming lectures on convex sets, will help proving it. Infact we mentioned all duality concepts including that of notion of subgradients for convex functions follow from this basic, fundamental, separation theorem.

4. For now, we assumed that the above conjecture is true and hence dual description of a closed cone is immediate:

   Theorem 9.0.2. *K is a closed cone if and only if it is intersection of halfspaces through the origin.*

   Hence, we take this as the dual definition/characterization of Conic sets.

5. Another important result in duality is:

   Theorem 9.0.3. *K is a polyhedral cone if and only if it has a finite dual description.*

This we proved later while characterizing polyhedra.

6. The following results about algebra with cones K_1, K_2 are true:
   (a) (Arbitrary) intersection of cones is a cone.
   (b) Union of cones need not be a cone. However, $CONIC(K_1 \cup K_2) = K_1 + K_2$.
   (c) (Any) linear combination of cones is a cone.
   (d) Cartesian product of cones is a cone, and $(K_1 \times K_2)^* = K_1^* \times K_2^*$.
   (e) Complement of a cone is never a cone.
   (f) $K_1 \subseteq K_2 \Rightarrow K_2^* \subseteq K_1^*$.
   (g) Milutin-Dubovitski lemma: $(K_1 \cap K_2)^* = K_1^* + K_2^*$, for closed cones K_1, K_2 whose sum is also closed[1].
7. Following topological results hold for cones:
   (a) Cones can be closed, open, neither, both.
   (b) Cones are unbounded.
   (c) Refer to exercise B.15 in Nemirovski [2005].
8. Refer sections B.1.4, B.2.6.B in Nemirovski [2005], section 2.6.1 in Boyd and Vandenberghe [2004], and optionally relevant parts in sections 2, 14 in Rockafellar [1996], for discussion on cones.

[1]Proposition B.2.3 in Nemirovski [2005].

Lecture 10

Convex sets: Polytopes

1. We say C is a convex set iff $x, y \in C, \lambda \in [0,1] \Rightarrow \lambda x + (1-\lambda)y \in C$ i.e., if two points are in the set, then the entire line segment induced by them is also in the set.

2. Motivated by above, we defined convex combination as linear combination with the restriction that the combining coefficients must be non-negative and must sum to unity.

3. We defined convex hull:

$$CONV(S) \equiv \left\{ \sum_{i=1}^{m} \lambda_i v_i \mid \lambda_i \in \mathbb{R}^+,\ v_i \in S\ \forall\ i = 1, \ldots, m, \sum_{i=1}^{m} \lambda_i = 1, m \in \mathbb{N} \right\},$$

   i.e., the set of all vectors which can be expressed as convex combinations of those in the set.

4. Using induction, it was simple to show that C is convex if and only if $C = CONV(C)$, which we took as the primal definition/characterization of Convex sets.

5. We looked at several examples including the Birkhoff polytope[1] in the matrix space. This motivated us to define a polytope: P is a polytope iff $\exists\ S \ni P = CONV(S),\ |S| \in \mathbb{N}$. We argued that the set of permutation matrices ($n!$ matrices) generates the Birkhoff polytope. The set of all matrices with every row having a one in exactly one column position (n^n matrices) generates the set of all Stochastic matrices.

6. We then defined an n-dimensional simplex as $CONV(S)$, where S is an affinely independent set of size $n+1$.

[1] https://en.wikipedia.org/wiki/Birkhoff_polytope.

7. We defined dimension of an set as that of its affine hull i.e., $dim(S) \equiv dim(AFF(S))$. This motivates a new definition for convex sets: sets that have all simplices (of the same dimension as the set) formed by points in the set. So convex sets are "made up" of basic polytopes ranging from a line-segment to a simplex.

8. Refer sections B.1.1-B.1.5 in Nemirovski [2005], sections 2.1-2.3 in Boyd and Vandenberghe [2004]. Optionally, sections 2,3 in Rockafellar [1996].

Lecture 11

Convex Sets: Polyhedra, Polar

1. We continued giving examples of convex sets:

   (a) Polyhedron is a special convex set that is an intersection of a finite number of half spaces (that need not pass through origin).

   (b) Shifted-Cones are sets of the form $K + \{a_0\}$, where K is a cone and $a_0 \in V$.

   (c) Generic convex sets like unit sphere/ball: $B = \{v \in V \mid \|v\| \leq 1\}$, unit p-norm ball $\{v \in \mathbb{R}^n \mid \|v\|_p \leq 1\}$, where $p \in [0, \infty]$, ellipse: $\mathcal{E}_M \equiv \left\{v \in \mathbb{R}^n \mid v^\top M v \leq 1\right\}$, where $M \succ 0$ (all centered at origin).

2. Motivated by polyhedra and intuition that all (closed) convex sets might be (not necessarily finite) intersections of half spaces, we generalized the notion of dual cones: given a set $S \subseteq V$, we define its polar as $S^\circ \equiv \{v \in V \mid \langle v, s\rangle \leq 1 \ \forall \ s \in S\}$.

3. For many sets we visualized who the polar would look like. In particular, it was easy to see that:

   (a) Polar of a cone is same as (negative of) dual cone. Polar of a linear set is same as its orthogonal complement.

   (b) Polar of a set is a convex set, even if the set is non-convex.

   (c) Polar of a set is a closed set, even if the set is not closed.

   (d) Polar of a set contains origin, even if the set does not.

   (e) $S^\circ = (CONV(S))^\circ$.

4. We then began proving the most important duality result for convex sets:

Theorem 11.0.1. *If C is a closed convex set containing origin, then $(C^\circ)^\circ = C$. As a consequence, $(K^*)^* = K$, whenever K is a closed cone. And, $\left(L^\perp\right)^\perp = L$, whenever L is a linear set.*

Proving $C \subseteq (C^\circ)^\circ$ was easy. The other way, proved in proposition B.2.2 in Nemirovski [2005], requires the so-called separation theorem that will be stated and proved in the next lecture.

5. We then covered definitions related to this theorem:

   (a) We say two sets $S_1, S_2 \subseteq V$ are strictly separated iff there exists a $w \in V \neq 0$, such that:

   $$\min_{s_1 \in S_1} \langle w, s_1 \rangle > \max_{s_2 \in S_2} \langle w, s_2 \rangle.$$

   Also, in this case, we say "w strictly separates S_1, S_2".

   (b) Given a set $S \subseteq V$ and $x_0 \in V$, we define projection of x_0 onto S, as any vector $\Pi_S(x_0)$ that satisfies:

   $$\Pi_S(x_0) \in \arg\min_{s \in S} ||x_0 - s||.$$

   We gave examples where the projection does not exist, where it exists but is not unique, and where it uniquely exists.

6. Refer section B.2.6 in Nemirovski [2005], and section 14 in Rockafellar [1996].

Lecture 12

Convex Set: Dual definition

1. We began by stating and proving[1] the separation theorem:

   Theorem 12.0.1. *Let C be a closed convex set and $x_0 \notin C$. Then*

   (a) $\Pi_C(x_0)$ *exists and is unique.*

   (b) $\langle x_0 - \Pi_C(x_0), x - \Pi_C(x_0)\rangle \leq 0 \ \forall \ x \in C$.

   As a consequence, $x_0 - \Pi_C(x_0)$ *strictly separates C and x_0.*

2. From theorem 11.0.1, it follows that:

   Theorem 12.0.2. *C is closed convex if and only if it is an intersection of half spaces (that need not pass through origin).*

   We take this as the dual definition/characterization of (closed) convex sets. The proof follows by shifting origin such that the set contains origin and applying theorem 11.0.1 and then shifting back the origin.

3. Using above results we were able to show the following interesting result, called as the (homogeneous) Farkas Lemma[2]:

   Lemma 12.0.3. *Consider the following system of inequalities:*

$$\begin{aligned} Ax &= b \\ x &\geq 0. \end{aligned} \tag{12.1}$$

---

[1]Some proofs, like this one, appear in previous offering's notes: `https://1drv.ms/b/s!Au6Zdrbq2x4phu1rCuc-ZBseLtnnuA`, `https://1drv.ms/b/s!Au6Zdrbq2x4pgc9YPLmTTUMOHwfemg`.

[2]Refer section B.2.4 in Nemirovski [2005]. Refer theorem 1.2.1 and exercises 1.2-1.4 for other such "theorems on Alternative".

The above system is solvable if and only if the following is not solvable:

$$A^\top y \geq 0$$
$$b^\top y < 0. \tag{12.2}$$

4. Motivated by separation theorem's proof, we defined the notion of a supporting hyperplane: Given a set $S \subseteq V$ and a point on the boundary, $x_0 \in \partial S$, we say that the hyperplane $\{x \in V \mid \langle w, x - x_0 \rangle = 0\}$ is a supporting hyperplane of S at x_0 iff $\langle w, x - x_0 \rangle \leq 0 \; \forall \; x \in S$.

5. We then desired to show that all closed convex sets have a supporting hyperplane at all boundary points. We argued that this will need defining two cones: the tangent and normal, which will be defined in the next lecture.

6. Read sections B.1.6, B.2.5 in Nemirovski [2005] and section 2.5 in Boyd and Vandenberghe [2004]. Optional reading: section 11 in Rockafellar [1996].

Lecture 13

Convex Sets: Supporting Hyperplane, Polyhedral Chracterization

1. We defined tangent cone[1] of a set S at a point $s_0 \in S$ as all those directions along which one can move from s_0 and stay inside S. Formally, $\mathcal{T}_S(s_0) \equiv \{h \in V \mid \exists t > 0 \ni x_0 + th \in S\}$.

2. After some examples, we easily showed that:

   Theorem 13.0.1. *For a convex set, tangent cone at any point is indeed a cone. Moreover,* $\mathcal{T}_S(s_0) = CONIC\left(\{s - s_0 \mid s \in S\}\right)$.

3. We then defined its dual cone as the normal cone[2]: $\mathcal{N}_S(s_0) \equiv (\mathcal{T}_S(s_0))^*$.

4. Since by definition of a boundary point, x_0, of a closed convex set, C, there is atleast one direction moving along which one cannot stay inside the set (for any small movement), it is clear that the tangent cone is not V. Hence the Normal cone cannot be $\{0\}$, and there consequently there exists a $w \neq 0 \in \mathcal{N}_C(x_0)$. By definition of Normal cone, it follows that $\{x \in V \mid \langle w, x - x_0 \rangle = 0\}$ is a supporting hyperplane of C at x_0. We summarize this as the following important theorem:

   Theorem 13.0.2. *Let C be a closed convex set and $x_0 \in \partial C$. Then there exists a supporting hyperplane for C at x_0.*

[1] Nemirovski [2005] calls this the radial cone.

[2] Boyd and Vandenberghe [2004] defines normal cone as the negative of the dual cone of the tangent cone.

5. We then proved that all polyhedra are polyhedral cones shifted by a polytope, known as Minkowski-Weyl theorem:

   Theorem 13.0.3. *A set $\mathcal{P}$ is polyhedral if and only if there exist finite sets K, C such that $\mathcal{P} = CONIC(K) + CONV(C)$.*

6. We proved this theorem by first showing that a cone is polyhedral if and only if it has finite dual description (refer theorem 4.5.1 in LAURITZEN [2009]), using the Fourier-Motzkin's algorithm (theorem 1.2.2 in LAURITZEN [2009]).

7. While proving above, we defined projection of a set S_1 onto S_2 as

$$\Pi_{S_2}(S_1) \equiv \left\{ s \in \arg\max_{s_2 \in S_2} ||s_1 - s_2|| \ \ni \ s_1 \in S_1 \right\}.$$

8. Refer sections B.2.5 in Nemirovski [2005]. Section 2.5 in Boyd and Vandenberghe [2004], and section 11 in Rockafellar [1996].

Lecture 14

Real-Valued Functions over Hilbert Spaces

1. We quickly wrapped up our discussion on convex sets by noting:

   (a) Following theorem gives a 1-d characterization for convex sets. The utility of this theorem was illustrated while showing that $\{x \mid x^\top Ax + b^\top x + c \leq 0\}$ is convex whenever $A \succeq 0$ (and this set is non-empty):

   Theorem 14.0.1. *A set C is convex if and only if intersection of C with any line is convex, whenever the intersection is non-empty.*

   (b) The following results about algebra with convex sets C_1, C_2 are true (refer section B.1.5 in Nemirovski [2005]):

      i. (Arbitrary) intersection of convex sets is a convex.
      ii. Union of convex sets need not be convex. However, $CONV(C_1 \cup C_2) = C_1 + C_2$.
      iii. (Any) linear combination of convex sets is a convex set.
      iv. Cartesian product of convex sets is a convex set.
      v. Consider an Affine mapping defined by $y = Ax + b \in \mathbb{R}^m, x \in \mathbb{R}^n$ where A is $m \times n$ and $b \in \mathbb{R}^m$.
         A. $C \subseteq \mathbb{R}^n$ is convex $\Rightarrow$ its image under the affine mapping, i.e., $\{y = Ax + b \mid x \in C\}$ is convex.
         B. $C \subseteq \mathbb{R}^m$ is convex $\Rightarrow$ its pre-image under the affine mapping, i.e., $\{x \mid Ax + b \in C\}$ is convex.
      vi. Complement of a convex set is never a convex set.

   (c) Following topological results hold for convex sets:

      i. Convex sets can be closed, open, neither, both.

ii. Convex sets can be bounded, unbounded.

iii. We defined relatively interior point x_0 of S iff $B_\epsilon(x_0) \cap AFF(S) \subseteq S$. The set of all relatively interior points are relative interior $rint(S)$. We argued that all convex sets have non-empty relative interior (as they contain simplices).

iv. Refer to section B.16 in Nemirovski [2005].

2. We then began study of the final ingredient of a MP, which is a real-valued function over a subset in a Hilbert space i.e., $f : V \mapsto \mathbb{R}_{ext}$[1]. We define domain of f as $dom(f) \equiv \{x \in V \mid -\infty < f(x) < \infty\}$.

3. We defined (and gave examples) of some special sets associated with functions:

   (a) $graph(f) \equiv \{(x, f(x)) \mid x \in dom(f)\}$. By definition this set lies in the direct sum of the Hilbert space in which the domain lies, and the space of reals.

   (b) $epi(f) \equiv \{(x, y) \mid x \in dom(f),\ f(x) \leq y\}$. By definition this set lies in the direct sum of the Hilbert space in which the domain lies, and the space of reals.

   (c) Level set of f at $t \in \mathbb{R}$: $\mathcal{L}_t(f) \equiv \{x \in dom(f) \mid f(x) \leq t\}$. By definition this set lies in the space same as the domain.

4. We then defined some topologically related concepts:

   (a) f is said to be closed iff its epigraph is a closed set.

   (b) f is said to be bounded above iff $\max_{x \in dom(f)} f(x) < \infty$. f is said to be bounded below iff $\min_{x \in dom(f)} f(x) > -\infty$.

   (c) f is said to be continuous at $x_0 \in dom(f)$ iff for every convergent sequence in the domain to it, $\{x_n \in dom(f)\} \rightarrow x_0$, we have that $\{f(x_n)\} \rightarrow f(x_0)$. f is said to be continuous (everywhere) iff it is continuous at every point in its domain.

   (d) f is said to be L-Lipschitz continuous (or simply L-conts) iff $x, y \in dom(f) \Rightarrow |f(x) - f(y)| \leq L\|x - y\|$. We showed that every L-conts function is continuous. However functions like the simple 1-d quadratic is continuous but not L-conts.

[1] We consider the extended reals as the co-domain because we already know that the objective could itself be defined as the value of an MP (like in Cascaded MPs), which could be $\pm\infty$.

(e) f is said to be differentiable at $x_0 \in int(dom(f))$ iff

$$\exists\ \nabla f(x_0) \in V \ \ni\ \lim_{x \to x_0} \frac{f(x) - f(x_0) - \langle \nabla f(x_0), x - x_0 \rangle}{||x - x_0||} = 0.$$

If such a $\nabla f(x_0)$ exists, then it will be unique and it is called as the gradient vector. It is a simple exercise to show that $\langle \nabla f(x_0), u \rangle = \lim_{h \to 0} \frac{f(x_0 + hu) - f(x_0)}{h} \equiv D_f(x_0; u)$, the directional derivative of f at x_0 in the direction[2] u. More specifically:

Theorem 14.0.2. *The i^{th} entry of $\nabla f(x)$ is $\frac{\partial f(x)}{\partial x_i}$.*

(f) $f : \mathbb{R}^n \mapsto \mathbb{R}$ is said to be twice-differentiable at $x_0 \in int(dom(f))$ iff $\exists\ \nabla f(x_0) \in \mathbb{R}^n\ \ \nabla^2 f(x_0) \in \mathbb{R}^{n \times n}\ \ni$

$$\lim_{x \to x_0} \frac{f(x) - f(x_0) - \nabla f(x_0)^\top (x - x_0) - \frac{1}{2} (x - x_0)^\top \nabla^2 f(x_0) (x - x_0)}{||x - x_0||^2} = 0.$$

If such a $\nabla^2 f(x_0)$ exists, then it will be unique and it is called as the Hessian matrix. A basic result in calculus says that:

Theorem 14.0.3. *The $(i,j)^{th}$ entry in $\nabla^2 f(x)$ is $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. Now define functions $g_{x_0 u}(t) \equiv f(x_0 + tu)$. Then, $\frac{\mathrm{d}^2 g_{x_0 u}(t)}{dt^2} = u^\top \nabla^2 f(x_0 + tu) u$.*

[2] `http://people.whitman.edu/~hundledr/courses/M225/Ch14/Example_DirectionalDeriv.pdf` provides an example where all directional derivatives exist but the function is NOT differentiable!

Lecture 15

Linear, Affine and Conic Functions

1. A function $f : L \subseteq V \mapsto \mathbb{R}$ is linear[1] iff L is a linear set, and $f(\sum_{i=1}^{n} \lambda_i x_i) = \sum_{i=1}^{n} \lambda_i f(x_i)\ \forall\ x_i \in L, \lambda_i \in \mathbb{R}, n \in \mathbb{N}$ i.e., Image of a linear combination of some points under the function is the same linear combination of images of those points. Basically, functions where linear intra-extrapolation is exact. We take this as the primal definition.

2. After giving some examples we noted the following important result that was very easy to prove:

   Theorem 15.0.1. *f is linear if and only if $graph(f)$ is a linear set (in direct sum space $\mathcal{V} \bigoplus \mathbb{R}$) with a dimensionality same[2] as that of $dom(f)$.*

   (a) We first showed f is linear if and only if graph(f) is a linear set. This was straight-forward to prove. The proof also showed that if graph of a function is linear, then the function must be of the form $f(x) = \langle w, x \rangle$ for some $w \in L$, which is itself a linear function.

   (b) We then noted that $dim(dom(f)) \leq dim(graph(f)) \leq dim(dom(f)) + 1$. Also, since $(x, y) \notin graph(f)$ whenever $y \neq f(x)$, the dimensionality of the linear set is not $dim(dom(f)) + 1$. Hence $dim(graph(f))$ must be $dim(dom(f))$.

3. From the above, the dual definition follows:

[1] For the extended real number counterpart, the definition reads like: A function $f : V \mapsto \mathbb{R}_{ext}$ is linear iff $f(\sum_{i=1}^{n} \lambda_i x_i) = \sum_{i=1}^{n} \lambda_i f(x_i)\ \forall\ x_i \in dom(f), \lambda_i \in \mathbb{R}, n \in \mathbb{N}$, and $dom(f)$ is a linear set. For linear functions, we follow the convention that $f(x) = \infty\ \forall\ x \notin dom(f)$.

[2] In the space of $\mathcal{L} \bigoplus \mathbb{R}$ the graph is a hyperplane through the origin. Here $\mathcal{L}$ is the space induced by L.

Theorem 15.0.2. ***Riesz representation theorem:*** *A function $f : L \mapsto \mathbb{R}$, where L is linear, is linear iff there exists*[3] *a $w \in L$ such that $f(x) = \langle w, x \rangle \ \forall \ x \in L$.*[4] *Moreover*[5]*, the space of linear functions on L, called the dual space, is equivalent to the space induced by L itself.*

4. Refer Section B.2.8 in Nemirovski [2005], sections B.2.1-B.2.3 in Nemirovski [2005] (these were not covered in lectures but very useful to know); relevant parts of sections 17,19,21 in Rockafellar [1996].

5. Once linear functions are studied, affine[6] functions (and analogous results) are immediate: A function $f : A \mapsto \mathbb{R}$ is affine iff A is affine and $f(\sum_{i=1}^{n} \lambda_i x_i) = \sum_{i=1}^{n} \lambda_i f(x_i) \ \forall \ x_i \in A, \lambda_i \in \mathbb{R} \ \ni \ \sum_{i=1}^{n} \lambda_i = 1, n \in \mathbb{N}$ i.e., Image of an affine combination of some points under the function is the same affine combination of images of those points. We take this as the primal definition. Needless to say, all linear functions are affine.

6. Again, we can show:

   Theorem 15.0.3. *f is affine if and only if $graph(f)$ is an affine set of dimensionality same as that of A. If L_A is the linear set associated with A, f is affine*[7] *if and only if there exists a $u \in L_A, b \in \mathbb{R}$ such that $f(x) = \langle u, x \rangle + b$. This is the dual definition.*

7. A function $f : K \mapsto \mathbb{R}$ is conic[8] iff K is a cone and $f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i) \ \forall \ x_i \in K, \lambda_i \geq 0, n \in \mathbb{N}$ i.e., Image of a conic combination of some points under the function under-estimates the same conic combination of images of those points. We take this as the primal definition. Needless to say, all linear functions are conic. We proved that all norms are conic functions.

8. It was easy to show that:

   Theorem 15.0.4. *f is conic if and only if $epi(f)$ is conic.*

[3]This statemen can also be alternatively proved using orthonormal basis for L.

[4]For the extended real number counterpart, the dual definition reads like: A function $f : V \mapsto \mathbb{R}_{ext}$ is linear iff (a) $L \equiv dom(f)$ is a linear set, (b) there exists a $w \in L$ such that $f(x) = \langle w, x \rangle \ \forall \ x \in dom(f)$, and $dom(f)$ is a linear set. For linear functions, we follow the convention that $f(x) = \infty \ \forall \ x \notin dom(f)$.

[5]This additional qualification is left as an exercise to be proven.

[6]For the extended real number counterpart, the definition reads like: A function $f : V \mapsto \mathbb{R}_{ext}$ is affine iff $f(\sum_{i=1}^{n} \lambda_i x_i) = \sum_{i=1}^{n} \lambda_i f(x_i) \ \forall \ x_i \in dom(f), \lambda_i \in \mathbb{R} \ \ni \ \sum_{i=1}^{n} \lambda_i = 1, n \in \mathbb{N}$, and $dom(f)$ is an affine set. For affine functions, we follow the convention that $f(x) = \infty \ \forall \ x \notin dom(f)$.

[7]For the extended real number counterpart, everything is the same with the additional convention that $f(x) = \infty \ \forall \ x \notin dom(f)$.

[8]For the extended real number counterpart, everything is the same with the additional convention that $f(x) = \infty \ \forall \ x \notin dom(f)$.

9. We gave many examples: all semi-norms are conic. We gave examples of conic functions that are not dened on entire V , those whose value can be negative etc.

Lecture 16

Dual Definition and Support Functions

1. We defined a huge family of functions: Support function of a set $C \subseteq V$, evaluated at $x \in V$, is defined as $\mathcal{S}_C(x) \equiv \max_{y \in C} \langle x, y \rangle$. It was easy to show that support function is always a conic function. Moreover, it is also easy to show that its a closed function (as its epigraph is defined by an intersection of halfspaces).

2. From the dual definition of closed cones, it was clear that:

   Theorem 16.0.1. *A function is closed conic if and only if it is a support function (for some set). In other words, a function is closed conic if and only if it is pointwise maximum of a set of linear minorants of it.*

   g is said to be a minorant of f iff $g(x) \leq f(x) \ \forall \ x \in V$. This theorem provides the dual definition for (closed) conic functions.

3. After providing many examples of support functions, we defined the support function of a unit-norm ball (centered at origin) as the dual norm:

$$\begin{aligned} ||x||_* \equiv \max_{y \in V} \quad & \langle x, y \rangle, \\ \text{s.t.} \quad & ||y|| \leq 1. \end{aligned}$$

   It was easy to show that dual norm is indeed as norm.

4. We then defined the dual function, f^*: a function whose epigraph is the dual cone of the epigraph of a given function[1], f. We noted examples of functions,

[1] Note that dual function can be defined for non-conic functions too!

whose dual function does not exist, by citing functions whose dual cone can never be a (valid) epigraph. Then we showed that:

Theorem 16.0.2. *Let f be a closed conic function whose dual function, f^*, exists. Then:*

$$\begin{aligned} f^*(x) = \max_{y \in V} \quad & \langle x, -y \rangle, \\ \text{s.t.} \quad & f(y) \leq 1. \end{aligned}$$

Moreover, $(f^*)^* = f$. For such functions, Theorem 16.0.1 is hence a corollary of this theorem i.e., Every closed conic function, f, is the support function of the set, $\{x \mid f^*(-x) \leq 1\}$, provided f^* exists.

5. The proof follows from that written for theorem 16.0.1 and the fact that $f(\lambda x) = \lambda f(x)$ if $\lambda \geq 0$.

6. Refer section 13 in Rockafellar [1996] for conic functions.

Lecture 17

Convex Functions

1. A function $f : C \mapsto \mathbb{R}$ is convex[1] iff C is convex and $f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y) \; \forall \; \lambda \in [0,1]$. Using mathematical induction we showed that:

   Theorem 17.0.1. *If $dom(f)$ is convex, then f is convex if and only if $f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i) \; \forall \; x_i \in A, \lambda_i \geq 0 \; \ni \; \sum_{i=1}^{n} \lambda_i = 1, n \in \mathbb{N}$. We take this as the primal definition.*

   Needless to say, all linear, affine, conic functions are convex.

2. We gave our first non-conic example of a convex function as $f(x) = ||x||^2$, where $||\cdot||$ is any valid norm (in some abstract space). It was easy to show this from the primal definition. Nevertheless, we soon realized we will need more definitions if we need to give more examples.

3. We noted the famous Jensen's inequality, from which many other fundamental inequalities can be derived[2]:

   Theorem 17.0.2. *If f is convex and X is a random variable such that $\mathbb{E}[f(X)] < \infty$, then: $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.*

   Note that the condition in Jensen's inequality with a discrete random variable taking finite values is same as the primal definition (Hence this inequality can be taken as a "Stochastic" definition for convex functions). We

[1] For the extended real number counterpart, everything is the same with the additional convention that $f(x) = \infty \; \forall \; x \notin dom(f)$.

[2] Refer section 3.1.9 in Boyd and Vandenberghe [2004]. See proof2 in `https://en.wikipedia.org/wiki/Jensen%27s_inequality#Proofs`.

mentioned that many fundamental inequalities like the (generalized) AM-GM, Holders etc., are a consequence of Jensens inequality (with the convex function[3] as $-\log(x)$).

4. Again, it was easy to show that:

   Theorem 17.0.3. *f is a convex function if and only if epi(f) is a convex set.*

5. We noted examples of convex functions whose epigraphs are not closed and those which are convex in the interior of their domains but not convex in the entire domain.

6. We named a special convex function: Indicator function of a set S evaluated at $x \in V$ is defined as $I_S(x) \equiv \begin{cases} 0 & \text{if } x \in S, \\ \infty & \text{else.} \end{cases}$. Needless to say, I_C, is convex if and only if C is convex.

7. We then generalized the notion of support function, which is nothing but a pointwise maximum of a set of linear functions, to the notion of Fenchel dual/Conjugate/Legendre Transformation, $f^{'}$, of (an arbitrary) function f:

$$f^{'}(x) \equiv \max_{y \in V} \langle x, y \rangle - f(y), \tag{17.1}$$

   which is nothing but pointwise maximum of a set of affine functions. Note that indeed conjugate generalizes support function: $I_C^{'} = S_C$. In other words, for (the restricted class of) indicator functions, the notion of conjugate is exactly same as that of Support function.

8. It was easy to show that conjugate of any function is closed convex.

9. We computed (analytical forms) for conjugates of some functions.

10. Sections 3.1.1, 3.1.7,3.1.8,3.1.9 in Boyd and Vandenberghe [2004]; C.1 in Nemirovski [2005]; relevant parts in section 4 in Rockafellar [1996].

[3] We will prove $-\log(x)$ is a convex function later.

Lecture 18

Convex Functions: Duality and Sub-gradients

1. Using the "epigraph" trick we wrote down a relationship between conjugate and dual function of closed conic functions: $f^{'}(x) = I_{-\mathcal{L}_1(f^*)}(x)$,. In particular, this shows that conjugate is not a generalization of the notion of dual function[1].

2. We then noted the following important duality theorem:

   Theorem 18.0.1. *If f is closed convex, then $\left(f^{'}\right)^{'} = f$.*

3. From the above theorem, the dual definition of (closed) convex function is immediate:

   Theorem 18.0.2. *f is closed convex if and only if it is conjugate of some functions. In other words a function is closed convex if and only if it is pointwise maximum of a set of affine minorants of it.*

4. We mentioned that global properties of a function turn out to be local properties of the conjugate and vice-versa. This is the key advantage of this duality relationship. For example, $f^{'}(0)$, which is a local property of conjugate is equal to $-\min_{y\in V} f(y)$, which is a global property of the original function. Similarly, look at theorem 20.0.4.

5. The notion of conjugate also gives the following inequality: $f(x) + f^{'}(y) \geq \langle x, y\rangle \ \forall \ x, y \in V$. This is called as the Fenchel's inequality. Again, many fundamental inequalities can be derived from this.

[1] Bonus marks for those who derive formulae for the generalization of dual function.

6. Then we began looking more closely at the vector(s) defining the tightest affine minorants (the supporting hyperplane). This lead to the following definition: A vector $v(x_0) \in V$ is said to be a sub-gradient of f at $x_0 \in dom(f)$ iff $f(x) \geq f(x_0) + \langle v(x_0), x - x_0 \rangle \; \forall \; x \in V$. This inequality is called as the sub-gradient inequality.

7. We noted examples where sub-gradient does not exist, exists but not unique, uniquely exists. The set of all sub-gradients of f at $x_0 \in dom(f)$ is known as the sub-differential set, $\partial f(x_0)$. f is said to be sub-differentiable at x_0 iff $\partial f(x_0) \neq \emptyset$. A function is sub-differentiable iff it is sub-differentiable at every point in its domain.

8. From theorem 13.0.2, one can show that:

   Theorem 18.0.3. *If f is convex, then it is sub-differentiable in the relative interior of its domain*[2]*.*

[2] Bonus marks for students who give an example of a convex function that is NOT sub-differentiable at a boundary point of its domain.

Lecture 19

Convex Functions: Sub-gradients

1. From the sub-gradient inequality, we computed the sub-gradients for various functions.

Lecture 20

Convex Functions: First-order Characterization

1. It is easy to show that:

   Theorem 20.0.1. *The sub-differential set is a convex set (whenever it is non-empty). The Sub-differential set at an interior point in the domain is bounded.*

2. We have the result:

   Theorem 20.0.2. *Let f be a convex function and $x_0 \in int(dom(f))$. f is differentiable at x_0 if and only if the gradient is the only sub-gradient at x_0 i.e., $\partial f(x_0) = \{\nabla f(x_0)\}$.*

   The proof for only if part is easy[1]: for any subgradient v (at x_0), we must have $\langle v, u\rangle \leq \frac{f(x_0+hu)-f(x_0)}{h}$. Because it is differentiable at x_0, talking limits on both sides of the inequality gives: $\langle v, u\rangle \leq D_f(x_0; u) = \langle \nabla f(x_0), u\rangle$. Since this is true for all u, we have $v = \nabla f(x_0)$. For proof of the if part, please refer theorem 25.1 in Rockafellar [1996].

3. We next noted that functions with open domain are convex if and only if they are sub-differentiable. For differentiable functions with open domain, convexity is same as gradient satisfying sub-gradient inequality. We write these observations as the following first-order characterization:

   Theorem 20.0.3. *Let f be a continuous function defined on a convex domain, then f is convex if and only if it is sub-differentiable in the*

[1] Prop. C.6.5 in Nemirovski [2005] provides an alternate proof.

domain's (relative) interior. Let g be a continuous function on a convex domain, and is differentiable in its (relative) interior, then g is convex if and only if the gradient is a sub-gradient in the (relative) interior.

4. From above arguments it is easy to show that:

   Theorem 20.0.4.

   $$\partial f(x_0) = \arg\max_{y \in V} \langle x_0, y \rangle - f^{'}(y).$$

5. Read sections C.6.3, C.6.2, C.2.2 in Nemirovski [2005]; section 3.3 in Boyd and Vandenberghe [2004], section 12, 26, 23 in Rockafellar [1996].

Lecture 21

Second-order Characterization

1. We noted the important result that helps one in computing a sub-gradient:

   Theorem 21.0.1. *Let $f = \max(f_1, \ldots, f_n)$. Let all f_i be convex, then f is also convex. Moreover, if $I_0 \subseteq \{1, \ldots, n\}$ is an index set such that $f(x_0) = f_i(x_0) \ \forall \ i \in I_0, \ f(x_0) > f_j(x_0) \ \forall \ j \in I_0^c$, then*

   $$\partial f(x_0) = CONV\left(\cup_{i \in I_0} \partial f_i(x_0)\right).$$

   Again, we easily proved RHS $\subseteq$ LHS and left the converse as bonus a exercise.

2. From sub-gradient inequality it follows that $-\partial f(x_0) \subseteq \mathcal{N}_{\mathcal{L}_{f(x_0)}(f)}(x_0)$. This further says[1], $-CONIC\left(\partial f(x_0)\right) \subseteq \mathcal{N}_{\mathcal{L}_{f(x_0)}(f)}(x_0)$. Assuming the converse is also true[2]:

   Theorem 21.0.2. *If f is a convex function, $x_0 \in int(dom(f))$, and $\partial f(x_0) \neq \{0\}$, then*

   $$-CONIC\left(\partial f(x_0)\right) = \mathcal{N}_{\mathcal{L}_{f(x_0)}(f)}(x_0),$$

   we gave examples of cases where a sub-gradient direction need not be an instantaneous ascent direction. However, it is clear that the gradient direction (if it exists) is always an instantaneous ascent direction. This is the fundamental reason behind why non-differential functions are more difficult to optimize, even in the convex regime.

[1] f is convex implies all level sets are convex, which implies, the normal cone is indeed a cone. Hence if a set (sub-differential set) is a subset of this cone, then the set's (sub-differential set's) conic hull must also belong to this cone (Normal cone).

[2] Bonus marks to the student who proves the converse

3. We then moved to second-order characterization and proved the following theorem:

   Theorem 21.0.3. *$f : (a,b) \mapsto \mathbb{R}$ is convex[3] if and only if $\frac{\mathrm{d}^2 f(t)}{dt^2} \geq 0 \ \forall \ t \in (a,b)$. Moreover, a continuous function $g : [a,b] \mapsto \mathbb{R}$ is convex if and only if $\frac{\mathrm{d}^2 g(t)}{dt^2} \geq 0 \ \forall \ t \in (a,b)$.*

4. From the definition of convex functions and the above, the following theorem is immediate:

   Theorem 21.0.4. *Let f be a function defined over a convex domain that is twice-differentiable in the interior of its domain and is continuous everywhere. For every $x_0, u \in C$, define the 1-d restriction $g_{x_0 u}$ given by: $g_{x_0 u}(t) \equiv f(x_0 + tu) \ \forall \ t \ni x_0 + tu \in C$. f is convex if and only if $\frac{\mathrm{d}^2 g_{x_0 u}(t)}{dt^2} \geq 0 \ \forall \ t \in int(dom(g_{x_0 u})), \ \forall \ x_0, u$.*

   In the lecture, we mentioned that the above turns out to be an "easy" definition for many example functions, especially the ones in complicated Hilbert spaces.

5. From the above and theorem 14.0.3, it follows that:

   Theorem 21.0.5. *A continuous function defined on a convex domain and that is twice-differentiable in the domain's interior is convex if and only if the Hessian is psd at any point in the domain's interior.*

6. Refer sections C.3, C.2.2 in Nemirovski [2005]; Chapter 3 and especially 3.1.3 and 3.1.4 in Boyd and Vandenberghe [2004], relevant parts in sections 23-25 in Rockafellar [1996].

[3] For the first statement in the theorem, "a" may be $-\infty$ and/or "b" may be ∞.

Lecture 22

Convex Programs

1. An MP (1.3) is said to be a Convex Program (CP) iff its objective, f, and all constraint functions, g_i, are convex. Needless to say, domain of a CP will hence be always convex.

2. After giving examples of CPs, we noted the following fundamental questions about CPs:

   (a) What are some sufficient conditions for CPs being bounded ? The answer we noted was:

   Theorem 22.0.1. *A CP is bounded whenever its feasibility set is bounded.*

   (b) What are some sufficient conditions for CPs being solvable ? The answer we noted was:

   Theorem 22.0.2. *A CP is solvable whenever its feasibility set is compact and its objective is continuous.*

   (c) What are some sufficient conditions for CPs being uniquely solvable ?

   (d) What (first order) conditions characteize optimality (of a candidate) ?

Lecture 23

First order Optimality conditions

1. In view of the third question, we defined strictly convex functions: A function $f : C \mapsto \mathbb{R}$ is strictly convex iff C is convex and $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) \ \forall \ \lambda \in (0,1)$. Needless to say, all strictly convex functions are convex. The following theorem was immediate:

   Theorem 23.0.1. *A CP is uniquely solvable whenever its feasibility set is compact, its objective is continuous, and strictly convex.*

2. We then defined Unconstrained Convex Programs: CPs whose domain is the entire set of vectors (that form a finite dimensional Hilbert space) and whose feasibility set is same as its domain. Equivalently, a CP whose domain is entire set of vectors and there are no constraints is an unconstrained CP i.e., CPs of the form $\min_{x \in V} f(x)$. The following theorem was easy to prove:

   Theorem 23.0.2. *Let f be a convex function, such that $dom(f) = V$. Then,*
   $$x^* \in \arg\min_{x \in V} f(x) \iff 0 \in \partial f(x^*).$$

3. Infact, we then generalized this to:

   Theorem 23.0.3. *Let (1.3) be a CP with a differential objective i.e., f is differentiable everywhere in $\mathcal{X}$. Then,*
   $$x^* \text{ is a solution to } (1.3) \iff \nabla f(x^*) \in \mathcal{N}_{\mathcal{F}}(x^*).$$

4. In the subsequent lecture we will write down simplified expressions for the normal cone of feasibility set for special classes of CPs and re-write the above theorem 23.0.3[1] appropriately.

[1] Bonus marks to students who generalize this theorem to the case where the objective is NOT differentiable.

Lecture 24

KKT conditions

1. We begin by defining Polyhedrally Constrained Convex Programs (PCCPs) as CPs with an open domain, and the constrained functions are all restricted to be affine i.e., CPs of the form:

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & f(x) \\ \text{s.t.} \quad & \langle a_i, x \rangle \leq b_i \ \forall\ i = 1, \dots, m, \end{aligned} \tag{24.1}$$

where $\mathcal{X}$ is open and f is convex.

2. The following theorem follows from theorem 23.0.3:

Theorem 24.0.1. *Let (24.1) be a CP with differentiable objective. Then, x^* is a solution to (24.1) if and only if there exists $\lambda^* \in \mathbb{R}^m$ such that:*

(a) $x^* \in \mathcal{X}$, $\langle a_i, x^* \rangle \leq b_i, \lambda_i^* \geq 0 \ \forall\ i = 1, \dots, m$ (feasibility conditions).

(b) $\lambda_i^* (\langle a_i, x^* \rangle - b_i) = 0 \ \forall\ i = 1, \dots, m$ (complementary slackness conditions).

(c) $\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* a_i = 0$ (gradient conditions).

3. We then defined a regular CP as a CP (1.3) with the domain restricted to be (convex) open, all functions (objective, constraint) restricted to be (convex) differentiable and the Slater's condition is satisfied. Slater's condition says that for each non-affine constraint, there must exist $x_i \in \mathcal{X} \ \ni \ g_i(x_i) < 0$. We then defined a KKT point (x^*, λ^*) as any pair x^*, λ^* that satisfy the following three (sets of) conditions:

(a) $x^* \in \mathcal{X}$, $g_i(x^*) \leq 0, \lambda_i^* \geq 0 \ \forall\ i = 1, \dots, m$ (feasibility conditions).

(b) $\lambda_i^* g_i(x^*) = 0 \ \forall\ i = 1, \dots, m$ (complementary slackness conditions).

(c) $\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) = 0$ (gradient conditions).

4. From theorem 23.0.3:

   Theorem 24.0.2. *x^* is a solution to a regular CP if and only if there exists $\lambda^* \in \mathbb{R}^m$ such that (x^*, λ^*) is a KKT point.*

Lecture 25

KKT conditions: Examples

1. We discussed an example where the KKT conditions can be used to derive analytical form for the solution of the optimization problem that arises in defining dual norm of (entrywise) p-norm.

Lecture 26

Lagrange Duality

1. Through examples, we noted that the advantages of KKT conditions are to i) arrive at optimal solutions analytically ii) get certificate of optimality for a given feasible solution iii) exhaustively list all optimal solutions iv) get analytical form (instead of actual) optimal solution v) obtain analytical expression for optimal value vi) compare optimal solutions, optimal values vii) Motivates numerical methods for solving and serves as stopping-criteria vii) Motivates dual and provides dual solution too, as we shall see shortly.

2. From our experience with notions of orthogonal complement, dual cone, polar, dual function, conjugate, we then noted desirable properties to define a "dual" MP:

   Convexity: We insist that dual of any MP must be a convex program. For e.g., polar of any set is a convex set, conjugate of any function is convex etc.

   Outer Description: We insist that dual of a min MP is a max MP such that value of the "primal" minimization MP at any feasible solution is less than that of the dual at any of its feasible solution. Then, the function values in primal will not "overlap" with those in dual. For e.g., vectors in orthogonal complement do not overlap with the set and provide an "outer description"! This is more formally called as principle of Weak Duality.

   (A)symmetry: We insist that the optimal value of primal, if its a convex program, is equal to that of its dual. This is modeled from facts like: polar of polar of a convex set (that is closed and has origin) is the original set, conjugate of conjugate is original function. If $P \geq D \geq D(D)$ and $D(D) = P$, then $P = D$. So, we insist that the primal and dual have same optimal value. This is formally called as principle of Strong Duality.

Inheritance: We insist that when we define a dual, we reuse some older notions of duality like conjugate, dual cone etc. Using these existing notions, one should be able write down a dual for a given MP.

3. We began by studying a particular dual, called Lagrange dual, that satisfies all above desirables.

4. Given an MP (1.3), henceforth referred to as the primal, we defined Lagrangian: $\mathcal{L}(x, \lambda) \equiv f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$. It's domain is $\mathcal{X} \times \mathbb{R}^m$. We call x as primal variables, and λ_i as Lagrange multipliers or Lagrange Dual variables or simply, dual variables. We then define the Lagrange dual function: $\underline{\mathcal{L}}(\lambda) \equiv \min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda)$. Finally, we define the Lagrange Dual Problem as:

$$
\begin{aligned}
\max_{\lambda \in \mathbb{R}^m} \quad & \underline{\mathcal{L}}(\lambda), \\
\text{s.t.} \quad & \lambda \geq 0.
\end{aligned}
\tag{26.1}
$$

5. It was an easy exercise to show:

Theorem 26.0.1. *Let P be the value of (an arbitrary, perhaps non-convex) MP given by (1.3), and D be that of it's Lagrange dual (26.1). Then, P (Weak Duality). Moreover, (26.1) is (always) a Convex Program.*

Lecture 27

Lagrange Dual: Strong Duality and examples

1. From KKT conditions, it follows that:

   Theorem 27.0.1. *Let (1.3) be a regular convex program that is solvable and the objective, constraint functions are differentiable. Then, value of (1.3) is equal to that of its Lagrange dual (26.1) i.e., Strong Duality holds*[1]*. Moreover, the dual is also solvable and* (x^*, λ^*) *is a KKT point for (1.3) if and only if* x^* *is a solution for (1.3) and* λ^* *is a solution for (26.1).*

   Infact, this theorem can be tightened: theorem D.2.2 in Nemirovski [2005]; however proof is more involved. We henceforth assume D.2.2 is true.

2. Our proof also clearly shows how the Lagrange dual function is infact a conjugate function. Hence all the 4 desirable properties are satisfied by the Lagrange Dual Problem.

3. We note that the following are the advantages of analyzing the Lagrange dual problem:

   (a) Irrespective of the space in which primal variables lie, the dual variables lie in Euclidean space! This makes it easy for writing (numerical) solvers for problems in *any* domain.

   (b) If primal is also in Euclidean space, say $\mathbb{R}^n$, then the number of variables and constraints in primal, dual are respectively: n, m, and m, m. Hence associated trade-offs apply in numerically solving them.

[1]Note that these are merely sufficient conditions for strong duality. There are striking examples of non-convex problems where strong duality holds (refer section 3.5 in Nemirovski [2005]).

(c) Dual typically problem an alternative view of the optimization problem that may lead to profound intuitions and/or efficient solvers. For e.g., the dual problem of minimizing distance between two polytopes happens to be that of maximally separating them[2].

(d) Even feasible solutions to the dual gives profound insights into the primal: For example, Value of dual at any dual feasible point is a lower bound on the primal. Moreover, at optimality, if a dual variable is non-zero, then the corresponding constraint in primal will be active (by complementary slackness).

(e) Leads to theorems on alternative as shown in Nemirovski [2005].

(f) Most importantly, dual gives a nice "convex approximation" (lower bound) to non-convex problems.

4. We then wrote down simplified forms for the Lagrange dual for special classes of Convex Programs:

5. We defined a Linear Program (LP) as a special CP of the form:

$$\begin{aligned} \min_{x \in V} \quad & \langle c, x \rangle, \\ \text{s.t.} \quad & \langle a_i, x \rangle \leq b_i \ \forall \ i = 1, \ldots, m. \end{aligned} \tag{27.1}$$

Theorem 27.0.2. *The Lagrange dual of (27.1) is:*

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & -b^\top \lambda, \\ \text{s.t.} \quad & \lambda \geq 0, \ \ c + \textstyle\sum_{i=1}^m \lambda_i a_i = 0. \end{aligned} \tag{27.2}$$

This again can be written as an LP. Hence self-duality holds for LPs.

6. We defined a Quadratic Program (QP) as an MP of the form:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \tfrac{1}{2} x^\top P x + q^\top x, \\ \text{s.t.} \quad & a_i^\top x \leq b_i \ \forall \ i = 1, \ldots, m. \end{aligned} \tag{27.3}$$

7. We noted that this will be a convex program if and only if $P \succeq 0$. Moreover, the objective will be strictly convex if and only if $P \succ 0$.

Theorem 27.0.3. *If $P \succ 0$, the Lagrange dual of (27.3) is:*

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & -\tfrac{1}{2}\lambda^\top A^\top P^{-1} A \lambda - \lambda^\top \left(A^\top P^{-1} q + b \right) - \tfrac{1}{2} q^\top P^{-1} q, \\ \text{s.t.} \quad & \lambda \geq 0. \end{aligned} \tag{27.4}$$

This again can be written as a convex QP, hence self-duality holds.

[2] Refer `http://www.robots.ox.ac.uk/~cvrg/bennett00duality.pdf` for details.

Lecture 28

Conic Programs and Duality

1. We derived the Lagrange dual problem for a general convex QP:

   Theorem 28.0.1. *If $P \succeq 0$, then the Lagrange dual of QP (27.3) is:*

   $$\begin{aligned} \max_{\lambda\in\mathbb{R}^m, t\in\mathbb{R}} \quad & t, \\ \text{s.t.} \quad & \begin{bmatrix} P & q + \sum_{i=1}^m \lambda_i a_i \\ q^\top + \sum_{i=1}^m \lambda_i a_i^\top & -2\left(\sum_{i=1}^m \lambda_i b_i + t\right) \end{bmatrix} \succeq 0, \lambda \geq 0. \end{aligned} \tag{28.1}$$

2. Motivated by the above (and generalizing LPs) we defined a Semi-Definite Program (SDP):

   $$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & c^\top x, \\ \text{s.t.} \quad & B - \textstyle\sum_{i=1}^n x_i A_i \succeq 0. \end{aligned} \tag{28.2}$$

   Here the matrices B, A_i are symmetric matrices of size m. The constraints of the form in SDP are known as Linear Matrix Inequalities (LMI).

3. We note that if all matrices B, A_i are diagonal matrices, then SDP is same as LP. Secondly, (27.4) can be written as an SDP.

4. SDP happen to be an enormous class of CPs, with huge number of applications. Section 3.2 in Nemirovski [2005] presents a host of sets that can be represented by LMIs.

5. Further generalizing SDPs, we defined Conic Programs:

   $$\begin{aligned} \min_{x\in V} \quad & \langle c, x\rangle_V, \\ \text{s.t.} \quad & b -_W l(x) \in K \subseteq W. \end{aligned} \tag{28.3}$$

Here V, W are vector sets from different vector spaces. For e.g., in SDPs, $V = \mathbb{R}^n$ and $W = \mathcal{S}_m$ (symmetric matrices). $a -_W b = a +_W (-1.b)$, where $+_W$ is the addition operator in W space. $\langle \cdot, \cdot \rangle_V$ is the inner-product in V space. $l : V \mapsto W$ is a linear function (definition same as with scalar valued functions). K is a closed cone.

6. It is easy to see that LPs, QPs, SDPs, can all be written in (28.3) form.

7. We defined conic dual of (28.3) as[1]:

$$\begin{aligned} \max_{y \in W} \quad & -\langle b, y \rangle_W, \\ \text{s.t.} \quad & l^\top(y) +_V c = 0, \; y \in K^* \subseteq W. \end{aligned} \tag{28.4}$$

8. Interestingly, in special cases of LP, QP etc., the Lagrange dual (when primal is written in (1.3) form) will match Conic Dual (when primal i written in (28.3) form).

   Theorem 28.0.2. *Conic dual is (always) convex and, the Value of (28.3) $\geq$ value of (28.4), even if K is arbitrary. If K is closed convex, their values are the same.*

   The proof follows from infimal convolution theorem and also highlights the optimality conditions for this case.

[1] Holds even in case K is an arbitrary set.

Bibliography

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

NIELS LAURITZEN. LECTURES ON CONVEX SETS. Available at https://www.fmf.uni-lj.si/~lavric/lauritzen.pdf, 2009.

A. Nemirovski. Lectures On Modern Convex Optimization. Available freely at www2.isye.gatech.edu/~nemirovs/Lect_ModConvOpt.pdf, 2005.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

Sheldon Axler. *Linear Algebra Done Right*. Springer-Verlag, 1997.