Convex Optimization (CS709)

Instructor: Saketh

Contents

Lecture 1

1.1 Key Topics

Review of Euclidean Spaces (E.S.) — linear combinations, inner products; Special subsets in E.S. — subspaces, affine sets, their primal/dual characterizations

1.2 Key Defns. & Results

Definitions: Vector space, Linear combination, Linear function, Conjugate space, Convergence of sequence of vectors, Sub-space, Span, Linear independence, Basis, Dimension, Orthogonal basis, Orthogonal complement, Affine set, Affine hull, Affine combination, Affine independence, Affine basis, Affine dimension, hyperplane

Results: Conjugate space of E.S. is itself, Inner and outer characterization of subspaces and affine sets, Existence and uniqueness of subspace associated with affine set

1.3 Summary

In order to understand Convex optimization Problems (CPs), it is important to understand the space in which the variables live i.e., Euclidean Spaces (ES), notion of convex sets in ES and convex functions defined over them. This lecture focuses on reviewing some concepts regarding ES.

ES [or any Vector Space (VS)] is equipped with two basic operations $+$ (vector addition), . (scalar multiplication) which lead to the notion linear combinations. Vector spaces are closed under (finite) linear combinations. Dot product or inner product ($<,>$) is another important operation in ES and leads to definitions of lengths (norms), distances etc. Interestingly, $<,>$ helps in characterizing real-valued (infact, any vector-valued) linear function defined on ES. One can show that there is a bijective mapping from the set of linear functions on $\mathbb{R}^n$ to $\mathbb{R}^n$ itself. Infact, the vector space of (real-valued) linear functions (called the conjugate space) is equivalent to that of $\mathbb{R}^n$. Moreover, generalizations of this result give that any linear $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ can be written as $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ (here, $\mathbf{A}$ is matrix of size $m \times n$).

We next focussed our discussion on some important subsets of $\mathbb{R}^n$: subspaces, affine sets. Subspaces are subsets which themselves form vector spaces along with the $+,.$ of $\mathbb{R}^n$. E.g., are lines, planes etc., through the origin. Affine sets are subsets obtained by shifting subspaces i.e., lines, planes not necessarily passing through origin. We characterized subspaces, affine sets are those closed under linear and affine combinations respectively. This immediately raised the question can we find the minimal set of vectors which when linearly/affinely combined give the original subspace/affine-set ? The answer was a basis/affine-basis which is a set of linear-independent/affinely-independent set of vectors whose span/affine-hull is the give subspace/affine-set. Such minimal sets are called as basis/affine-basis and the size of which is called dimension/affine-dimension. Such an analysis gives an "inner" (or builder's or primal) description of these sets i.e., description in terms of elements of the subsets. This leads to the question can we provide an "outer" (or sculptor's or dual) description ?

The answer is again positive: subspaces are always solutions of a finite system of homogeneous equations $\mathbf{A}\mathbf{x} = \mathbf{0}$, where rows of $\mathbf{A}$ are vectors in the orthogonal complement of the given subspace (and hence an outer description). Needless to say, the minimal outer description is obtained by using basis vectors of the orthogonal complement. Similarly affine-sets are always solutions of a finite system of non-homogeneous equations $\mathbf{A}\mathbf{x} = \mathbf{b}$. It is easy to come-up with examples where inner description is more efficient representation than the outer and vice-versa.

1.4 Recommended Reading

- Section A.1 to A.4 in Nemirovski [2005]
- Part 1 — section 1 in Rockafellar [1996]
- Linear algebra notes at `http://www.cse.iitb.ac.in/saketh/teaching/cs723Scans3.pdf`

Lecture 2

2.1 Key Topics

Definition and examples of convex sets, Inner characterization, operations preserving convexity

2.2 Key Defns. & Results

Definitions: Convex set, Halfspace, Polyhedra, Normed ball (Ellipsoid), Normed cone (Second Order Cone), ϵ-Neighbourhood of a set, Convex combination, Convex hull, Polytope, Simplex, Extreme points

Results: Characterization of convex sets using convex combinations, Convexity is preserved under Affine transformations

2.3 Summary

In this lecture we identified another important class of subsets of $\mathbb{R}^n$, the convex sets. Following the definition, many examples of convex sets were illustrated. In particular we saw that every Halfspace, Polyhedra, Normed ball, Normed cone, ϵ-Neighbourhood of a convex set, Polytope (hence Simplex) are examples of convex sets. We also argued out why arbitrary intersections of convex sets are convex (union of convex sets need not be convex) and proved that convexity of sets is preserved under generic affine transformations. As with subspaces and affine sets, we worked towards inner (and later on outer) descriptions of convex sets and accordingly showed that all convex sets are closed under (finite) convex combinations of constituent vectors. We then defined convex hull of a set as that containing all convex combinations of vectors in the given set. It was immediate to see that a notion of extreme points ("basis" for convex set) exists.

2.4 Recommended Reading

- Sections B.1.1 to B.1.3 and B.1.5 in Nemirovski [2005]
- Part 1 — section 2, section 3 in Rockafellar [1996]
- Sections 2.1 – 2.2.4, 2.3 in Boyd and Vandenberghe [2004]

Lecture 3

3.1 Key Topics

Further operations preserving convexity, Conic sets, Summary of inner/outer characterization of all classes of sets, Illustration of special subsets in spaces of matrices, Topology of sets

3.2 Key Defns. & Results

Definitions: Sum (scaling) of two (or more) sets, Linear fractional function, Conic set (Cone), Conic combination, Polyhedral cone, Frobenius inner product (and consequently Frobenius norm), Positive-semidefinite cone, Closed set, Open set, Closure of a set

Results: $S_i, i \in \Lambda$ is a family of convex sets and $\lambda_i \in \mathbb{R}, i \in \Lambda$, then $\sum_{i\in\Lambda} \lambda_i S_i$ is convex, Image of convex set under Linear fractional transformation is convex

3.3 Summary

We studied two more important operations that preserve convexity (see Key results above). We formally defined Conic sets as those closed under conic combinations and studied few examples. We noted that Normed cones are only a special kind of cones. Also, the following important table was discussed:

Exploiting the fact that Matrix spaces (vec. space with vec. as matrix) are equivalent to Euclidean spaces (once matrix is expanded column/row-wise and using Frobenius inner product), the special class of subsets of this new space were studied using various examples. Some examples of subspaces were set of all

Table 3.1: Special subsets of $\mathbb{R}^n$

Class of subset	**Defn.**	**Inner Desc.** $\sum_i \lambda_i \mathbf{v}_i$	**Outer Desc.**	**E.g.**
Subspace	Closed under lin. comb.	lin. comb. of Basis vec. ($\lambda_i \in \mathbb{R}$)	$\{\mathbf{x} \mid \mathbf{Ax}=\mathbf{0}\}$	lines, planes thru origin
Affine set	Shifted subspaces	aff. comb. of aff.-basis vec. ($\sum_i \lambda_i = 1$)	$\{\mathbf{x} \mid \mathbf{Ax}=\mathbf{b}\}$	lines, planes
Cone	Closed under conic comb.	conic comb. of extreme pts. ($\lambda_i \geq 0$)	(closed) $\{\mathbf{x} \mid \mathbf{a}_\alpha^\top \mathbf{x} \leq 0, \alpha \in \Lambda\}$	polyhedral, normed cones
Convex set	Contains all line seg.	convex comb. of extreme pts. ($\lambda_i \geq 0, \sum_i \lambda_i = 1$)	(closed) $\{\mathbf{x} \mid \mathbf{a}_\alpha^\top \mathbf{x} \leq b_\alpha, \alpha \in \Lambda\}$	polyhedra, polytope, ball

symmetric/upper-triangular/lower-triangular/diagonal matrices. Some examples of affine-sets were all matrices with lower-part equal to unity etc.,and in general $\{\mathbf{X} \mid trace(\mathbf{A}_i, \mathbf{X}) = b_i, i = 1, \ldots, n\}$. Classic e.g., of conic set is the semi-definite cone and that of a convex set is normed ball $\|\mathbf{X}\| \leq 1$ where $\|\cdot\|$ is any matrix norm.

In view of studying topological properties of convex sets, we began defining things like open set, closed set, closure of a set and gave many examples.

3.4 Recommended Reading

- Sections A.7, B.1.4 in Nemirovski [2005]
- Sections 2.2.5, 2.3.3 in Boyd and Vandenberghe [2004]

Lecture 4

4.1 Key Topics

Notions of interior & relative interior, Interesting topological properties of convex sets, 3 fundamental theorems on convex sets

4.2 Key Defns. & Results

Definitions: Interior, Relative interior

Results: Complementation of open set is closed (vice-versa), rel.int. of non-empty convex set is non-empty, rel.int. and closure of convex set is convex, If S is convex then $Cl(ri(S)) = Cl(S)$ and $ri(Cl(S)) = ri(S)$, Caratheodary and Radon and Helly theorems.

4.3 Summary

We noted the following results regarding closed/open sets: arbitrary intersection of closed sets is closed (arbitrary intersection of open sets need not be open), arbitrary union of open sets is open (arbitrary union of closed sets need not be closed), complement of open set is closed, and that of closed set is open. Notions of interior and relative interior were introduced with illustrative examples. In general, $ri(S) \subset S \subset Cl(S)$. However for generic sets, the relations might be extremely loose (e.g., set of rationals). Convex sets have very special topological properties: rel.int of non-empty convex set is non-empty (convex sets are fat not thin) and infact the closure can be recovered from radial closure of the rel.int., in other words, the rel.int. is considerable.

The key concept in many of the theorems we studied was that a convex set will have simplices sitting in them. See above section on results for a complete list. We then noted three fundamental theorem about convex set which are all equivalent and again use of the notion of simplex: Caratheodary, Radon and Helly theorems. The statements of the theorems were studied. It is indeed useful to go through their formal proofs.

4.4 Recommended Reading

- Sections B.1.6, B.2.1, B.2.2, B.2.3 in Nemirovski [2005]
- Part-2 section 6; Part-4 section 17,21 in Rockafellar [1996]

Lecture 5

5.1 Key Topics

Notion of various kinds of Separability, Separability theorems; Farkas lemma, Theorems on Alternative; Outer description of closed convex sets

5.2 Key Defns. & Results

Definitions: Proper and strict/strong separation, GLB (infimum), LUB (supremum), boundedness, compactness, distance between sets, projection of a point onto a convex set, supporting hyperplane

Results: Characterization of properly(and strictly) separable convex sets, Farkas lemma, Outer Description of closed convex sets

5.3 Summary

After defining proper, strict separability of sets in a compact fashion (using sup,inf etc.), we went on characterize criteria for proper/strict separability in case of convex sets:

Theorem 5.3.1. *Given that S_1, S_2 are convex, S_1, S_2 are properly separable if and only if $ri(S_1) \cap ri(S_2) = \phi$.*

Theorem 5.3.2. *Given that S_1, S_2 are convex, S_1, S_2 are strictly/strongly separable if and only if $d(S_1, S_2) > 0$.*

We briefed on the proof of theorem 5.3.1: the "only if" part was easy (proof by contradiction). Proof of the "if" part was based on following theorems:

Theorem 5.3.3. ***Projection Theorem:*** *If S is closed and convex and $x \notin S$, then:*

1. *Projection of x onto S i.e., $P_S(x) = \text{argmin}_{y \in S} ||x-y||$ exists and is unique.*
2. *$y_0 = P_S(x)$ if and only if for every $y \in S$, we have $(y - y_0)^\top (y_0 - x) \geq 0$*

Immediate consequence of this theorem is (proof by construction):

Theorem 5.3.4. *If S is closed convex and $x \notin S$, then S and $\{x\}$ are strictly separable*

The following important deep result follows from above theorem:

Lemma 5.3.5. ***Farkas Lemma:*** *Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{b} \in \mathbb{R}^m$. Then the following statements are equivalent:*

1. *$\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq 0$ has solution*
2. *$\mathbf{A}^\top \mathbf{y} \geq 0, \mathbf{b}^\top \mathbf{y} < 0$ has no solution*

Farkas lemma and its generalizations (theorems on alternative) lead to notions of duality. From Farkas lemma one can derive the following:

Theorem 5.3.6. *If S is closed convex and $x \notin S$, then S and $\{x\}$ are properly separable.*

Once this is in place, it is easy to prove the 'if' part of theorem 5.3.1 (using the trick of creating the difference set).

The separability theorems are very useful and we illustrated one application — every convex set can be represented as: $\{\mathbf{x} \mid \mathbf{a}_\alpha^\top \mathbf{x} \leq b_\alpha, \ \alpha \in \Lambda\}$ — intersection of (infinite) halfspaces. This naturally lead to the definition of supporting hyperplane.

5.4 Recommended Reading

- Section B.2.5 in Nemirovski [2005]
- Part-3 section 11 in Rockafellar [1996]

Lecture 6

6.1 Key Topics

Definition and alternate characterizations of convex functions, examples.

6.2 Key Defns. & Results

Definitions: Convex function, Strict convex function, Epi-graph, Level-set, Quasi-convex functions, Directional-derivative, Gradient

Results: Epi-graph and first-order characterization of convex functions, Level-sets of convex functions are convex

6.3 Summary

We began with the definition of convex(concave) functions and strictly-convex functions. We defined the epi-graph of convex function and showed that it will be a convex set and also the converse, hence leading to a characterization of convex functions as exactly those having convex epi-graphs. We also defined level-sets, and showed that level-sets for a convex function are convex sets. However not all functions with convex level-sets are convex (eg. Gaussian function); such functions which have convex level-sets are called as quasi-convex functions.

We noted that epi-graph of convex function is not only convex but also closed; hence has supporting hyperplane at every point. This motivated the first-order characterization of convex functions:

Theorem 6.3.1. *Given f is differentiable, then f is convex if and only of* $f(\mathbf{x}) \geq f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x} - \mathbf{x}_0)\ \forall \mathbf{x}, \mathbf{x}_0$

Prior to proving this theorem, we revised the notions of directional derivative, gradient and differentiability of real-valued functions on $\mathbb{R}^n$. We noted that the gradient gives the direction of maximum increase of the function and hence is always Normal to the corresponding contour/level-set of f. From the above theorem, it is also easy to see that the supporting hyperplane at $\mathbf{x}_0$ is described as that plane orthogonal to the vector $[\nabla f(\mathbf{x}_0) \quad -1]^\top$ and passing through $(\mathbf{x}_0, f(\mathbf{x}_0))$. Hence the gradient is also useful in describing the supporting hyperplane of a convex function. The proof of theorem 6.3.1 is based on extending the same result in case of real-valued functions on $\mathbb{R}$ and using the following necessary condition for convexity of functions:

Lemma 6.3.2. *If $f : S \mapsto \mathbb{R}$ is convex (here $S \subset \mathbb{R}^n$ is a convex set) then $g : [0,1] \mapsto \mathbb{R}$, defined by $g(t) \equiv f(t\mathbf{x} + (1-t)\mathbf{y})$, is convex for all $\mathbf{x}, \mathbf{y} \in S$.*

6.4 Recommended Reading

- Sections 3.1.1-3.1.7 in Boyd and Vandenberghe [2004]
- Preliminaries in sections 4, 25, 23 in Rockafellar [1996]
- Sections C.1, C.3 in Nemirovski [2005]

Lecture 7

7.1 Key Topics

Jensen's inequality, Necessary and sufficient conditions for local(global) minimizer of a convex function (on entire $\mathbb{R}^n$), Second-order characterization of convex functions

7.2 Key Defns. & Results

Definitions: Global/local minimum/minimizer, Hessian, positive semi-definite (psd) matrix

Results: Jensens' inequality — AM-GM, Holder's inequalities; For a convex and differentiable f, x^* is a global minimizer if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$, A continuously twice differentiable f is convex if and only if the Hessian at every point is psd.

7.3 Summary

We started with a generalization of the definition of convex functions leading to the very useful Jensen's inequality. This inequality leads to various fundamental inequalities like the AM-GM, Holder inequalities. The proof of Jensen's inequality was based on the convexity of the epi-graph of f.

We recalled the definitions of local/global minimum/minimizer and the analysis result that $\nabla f(\mathbf{x}^*) = 0$ is necessary condition for $\mathbf{x}^*$ to be a local minimizer for any differentiable f (need not be convex). For convex functions, this necessary condition turned out to be sufficient — this followed from the first-order characterization of convex functions. We note some useful corollaries like: i) every local minimizer is a global minimizer for a convex function ii) set of global/local minimizers for a convex function is itself convex — this is because the level sets are convex iii) strictly convex functions either do not have a minimizer or have a unique one.

For real-valued functions on $\mathbb{R}$, we then came up with yet another characterization ($g''(x) \geq 0 \ \forall \ x$). Using the fact that 1-d restrictions of convex f are themselves convex and the notions of Hessian, we then extended the characterization to multivariate f as exactly those with Hessian being psd at all $\mathbf{x}$.

Using the 4 characterizations of convex functions learnt till now, we proved convexity of the some of the well known functions.

7.4 Recommended Reading

- Sections 3.1.4-3.1.5, 3.1.8-3.1.9 in Boyd and Vandenberghe [2004]
- Theorem 4.5 Rockafellar [1996]
- Sections C.1.2.A, C.5 Nemirovski [2005]

Lecture 8

8.1 Key Topics

Operations preserving convexity, Properties of Convex Functions

8.2 Key Defns. & Results

Definitions: Linear combination of functions (point-wise summation, scaling), point-wise maximum/supremum of functions, Bounded function and Locally bounded function, Lipschitz continuity and local Lipschitz continuity.

Results: i) conic combinations of a (uncountable) set of convex functions is a convex function (and hence in vector space of functions, convex functions form a cone), ii) point-wise supremum of convex functions is convex, iii) composition $h = f(g)$ of monotonically non-decreasing(increasing) f and convex(concave) g is convex. iv) Convex functions are locally bounded and locally Lipschitz continuous.

8.3 Summary

We began with discussion of operations preserving convexity and derived results i)-iii) above. Knowledge of such operations might be helpful in ascertaining convexity of functions that you might encounter in your research.

Convex functions are special ones in many respects and we attempted to understand few properties satisfied by them. We first proved that convex functions are locally bounded. Though locally bounded from both sides, it is the boundedness from below is what is obviously interesting (as we specifically assume that $dom(f) = \{\mathrm{x} \mid f(\mathrm{x}) < \infty\}$, boundedness from above is not surprising). We used this result to show that convex functions are locally Lipschitz continuous.

8.4 Recommended Reading

- Sections 3.2.1-3.2.4 in Boyd and Vandenberghe [2004]
- Section C.4 in Nemirovski [2005]

Lecture 9

9.1 Key Topics

Concept of sub-gradients and sub-differential, Characterization of closed convex functions as exactly those with non-empty sub-differential in rel-interior, Necessary and Sufficient conditions for minimizer of convex functions on $\mathbb{R}^n$, Review of Farka's Lemma and Theorems on Alternative.

9.2 Key Defns. & Results

Definitions: Sub-gradient, Sub-differential, Sub-differentiability

Results: i) If function has non-empty sub-differential everywhere on the entire (convex) domain, then it has to be a convex function ii) A closed convex function f has non-empty sub-differential everywhere on $ri(dom(f))$ iii) Sub-differential at an rel-interior point, for convex functions, is a compact convex set iv) x^* is a global/local minimizer of f (a convex function on $\mathbb{R}^n$) if and only if $0 \in \partial f(x^*)$.

9.3 Summary

Motivated by the first-order characterization of differentiable convex functions, and the fact that support hyperplanes exist for convex functions with closed epi-graphs (closed convex functions), we defined the notion of sub-gradient (sub-differential etc.). We then proved interesting results i) and ii) which give us a first-order kind of characterization for non-differential (closed) convex functions. Proof of i) was trivial; whereas ii) involved showing that the supporting hyperplane is never vertical — which followed from Lipshcitz continuity. Then we proved iii) which in particular showed the interesting result that the sub-gradient is always bounded by the (local) Lipshcitz constant. We then proved iv) (which was trivial) and concludes the discussion about un-constrained convex optimizations problems.

Towards the end of the lecture we refreshed results like Farka's Lemma and theorems on alternative which, as we will see in the next lecture, naturally lead to the concept of duality!

9.4 Recommended Reading

- Sections 3.1.3, 3.1.5, 3.1.6 in Nesterov [2004]
- Section C.6 in Nemirovski [2005]
- Scanned pages on Theorems on Alternative.

Lecture 10

10.1 Key Topics

Duality in linear programs.

10.2 Key Defns. & Results

Definitions: Mathematical Program (MP); Domain, objective function, constraint functions of an MP; Convex Program (CP); Linear Program (LP); feasible solution, feasible set, feasibility of a MP; optimal value, boundedness of an MP; optimal solution, optimal set, solvability of a MP;

Results: Notion of dual (in LPs) and the following theorem:

Theorem 10.2.1. *Let* $\boldsymbol{P}$ *be the primal LP and* $\boldsymbol{D}$ *its dual mentioned in Lecture:*

$$\begin{array}{lll} \mathbf{P} & \max\limits_{\mathbf{x}} & \mathbf{c}^\top \mathbf{x} \\ & \text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{b} \end{array}$$

$$\begin{array}{lll} \mathbf{D} & \min\limits_{\mathbf{y}} & \mathbf{b}^\top \mathbf{y} \\ & \text{s.t.} & \mathbf{A}^\top \mathbf{y} = \mathbf{c},\ \mathbf{y} \geq 0 \end{array}$$

Then, the following statements are true:

1. $\boldsymbol{P}$ *is bounded if and only if* $\boldsymbol{D}$ *is solvable*

2. $\boldsymbol{D}$ *is bounded if and only if* $\boldsymbol{P}$ *is solvable*

3. *If* $\mathbf{x}^*$ *is an optimal solution of* $\boldsymbol{P}$ *and is* $\mathbf{y}^*$ *is an optimal solution of* $\boldsymbol{D}$*, then* $\mathbf{c}^\top \mathbf{x}^* = \mathbf{b}^\top \mathbf{y}^*$ *(zero-duality gap cond.) and* $(\mathbf{A}\mathbf{x}^* - \mathbf{b})^\top \mathbf{y}^* = 0$ *(complementary slackness cond.)*

*4. If **P** is unbounded then **D** is infeasible and if **D** is unbounded then **P** is infeasible.*

10.3 Summary

The lecture began with definitions of the terms in sec. 10.2. Subsequently, for the problem $\mathbf{P}$, a notion of dual was sought for. The idea was to re-write the fact that $\mathbf{P}$ has optimal value α^* in terms of feasibility of this system of linear inequalities:

$$\mathbf{S1} \qquad \begin{aligned} \mathbf{c}^\top \mathbf{x} &> \alpha \\ \mathbf{A}\mathbf{x} &\leq \mathbf{b} \end{aligned}$$

Note that the statement "$\mathbf{S1}$ is infeasible for $\alpha \geq \alpha^*$ and feasible for $\alpha < \alpha^*$" is exactly same as saying "$\mathbf{P}$ has optimal value α^*". Once this is in place, we represent $\mathbf{S1}$ with an equivalent "dual/outer" system $\mathbf{S2}$ using theorem on alternative (extension of Farka's Lemma):

$$\mathbf{S2} \qquad \begin{aligned} \mathbf{b}^\top \mathbf{y} &<= \alpha \\ \mathbf{A}^\top \mathbf{y} &= \mathbf{c}, \mathbf{y} \geq 0 \end{aligned}$$

Note that $\mathbf{S2}$ is feasible whenever $\mathbf{S1}$ is infeasible and vice-versa; moreover the statement "$\mathbf{S2}$ is feasible for $\alpha \geq \alpha^*$ and in-feasible for $\alpha < \alpha^*$" is exactly same as saying "$\mathbf{D}$ has optimal value α^*". In other words, $\mathbf{P}$ and $\mathbf{D}$ are seemingly two different optimization problems, with different characteristics, however solving either of them gives the optimal value α^*. We will see this equivalence turns out to be very useful in many places. Once this equivalence is proved, we went through a proof for the above important theorem.

10.4 Recommended Reading

- Sections 1.1, 1.2 in Nemirovski [2005]

Lecture 11

11.1 Key Topics

Duality in convex programs; Characterization of solution in terms of saddle point

11.2 Key Defns. & Results

Definitions: Lagrangian function, Lagrange multipliers, weak duality, strong duality, Slater's condition, Regular Convex Programs (RCPs), saddle point

Results: The following key theorem:

Theorem 11.2.1. *Let **P** be a given MP and **D** its dual mentioned in Lecture:*

$$\mathbf{P} \quad \begin{array}{rl} \min\limits_{\mathrm{x}\in X} & f(\mathrm{x}) \\ \text{s.t.} & g_i(\mathrm{x}) \leq 0, \ \ i = 1, \ldots, m \end{array}$$

$$\mathbf{D} \quad \begin{array}{rl} \max\limits_{\lambda} & \underline{L}(\lambda) \\ \text{s.t.} & \lambda \geq 0 \end{array}$$

where $\underline{L}(\lambda) = \inf_{\mathrm{x}\in X} L(\mathrm{x}, \lambda)$ *and* $L(\mathrm{x}, \lambda) = f(\mathrm{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathrm{x})$ *is the Lagrangian function (λs are called the Lagrange multipliers). Then, the following statements are true:*

1. *Optimal value of **P** $\geq$ that of **D** (weak-duality)*
2. *For RCPs, optimal value of **P** $=$ that of **D** (strong-duality, proved using separation theorem)*

3. *For RCPs, **P** is bounded $\Rightarrow$ **D** is solvable*

4. *Objective of **P** is equal to* $\inf_{\mathrm{x}\in X} \bar{L}(\mathrm{x})$ *where* $\bar{L}(\mathrm{x}) = \sup_{\lambda\geq 0} L(\mathrm{x}, \lambda)$.

5. *$\underline{L}$ is always a concave function; whereas for RCPs $\bar{L}$ is convex.*

6. *If $(\mathrm{x}^*, \lambda^*)$ is a saddle point of L, then x^* is a solution of **P** and λ^* is a solution of **D** and more importantly, optimal values of **P** and **D** are equal.*

7. *For RCPs, if x^* is a solution of **P**, then $\exists\lambda^* \geq 0 \ni (\mathrm{x}^*, \lambda^*)$ is a saddle point of L.*

11.3 Summary

The idea was to extend the learning in case of duality of LPs to the case of generic programs. We repeated the process of writing the fact that **P** has a certain optimum value in terms of feasibility of a system and then looking at a dual system's feasibility problem and hence writing a corresponding dual problem. In particular, we proved the above key theorem. At the heart of which is the proof utilizing the separation theorem (proof of strong-duality). The main take-homes from this theorem are that: i) for RCPs, strong-duality holds and infact solvability of **D** is assured as long as **P** is bounded ii) for RCPs, given $\mathrm{x}^* \in X, \lambda^* \geq 0$, x^* is solves **P** if and only if $(\mathrm{x}^*, \lambda^*)$ is a saddle point of **P**. More importantly from the proof of above theorem, the following is clear: a saddle point for the Lagrangian exists if and only if **P**,**D** are solvable and strong-duality holds.

Results obtained for Lagrangian function above, motivate the study of generic problems involving function of two sets of variables, and quantity of interest is maximization over one set while minimizing over the other (min-max problems). We noted the following useful theorem (without proof), known as Sion-Kakutani theorem, in the lecture:

Theorem 11.3.1. *Let X, Y be two convex and compact sets in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Let $L : X \times Y \mapsto \mathbb{R}$. Now define two functions:* $\bar{L}(\mathrm{x}) = \sup_{\mathrm{y}\in Y} L(\mathrm{x}, \mathrm{y})$ *and* $\underline{L}(\mathrm{y}) = \inf_{\mathrm{x}\in X} L(\mathrm{x}, \mathrm{y})$. *Consider the problems:* $\min_{\mathrm{x}\in X} \bar{L}(\mathrm{x})$, *denoted as problem **I** and* $\max_{\mathrm{y}\in Y} \underline{L}(\mathrm{y})$, *denoted as problem **II**. If L is a continuous function and is convex on $\mathrm{x} \in X$ for every fixed $\mathrm{y} \in Y$ and is concave on $\mathrm{y} \in Y$ for every fixed $\mathrm{x} \in X$, then:*

- ***I** and **II** are solvable.*

- *Optimal values of **I** and **II** are the same.*

(Infact, for any L, the above two properties are necessary and sufficient for existence of saddle points for L)

In particular, the theorem gives a scenario where the min and max can be exchanged!

11.4 Recommended Reading

- Whole of appendix section D, except pages 415-416 in Nemirovski [2005]
- Section 5 upto subsection 5.2.4 (excluding 5.1.6) and including sections 5.3.2, 5.4 in Boyd and Vandenberghe [2004]

Lecture 12

12.1 Key Topics

An example where strong duality holds in non-convex programs, Optimality conditions for (constrained, regular) convex programs

12.2 Key Defns. & Results

Definitions: Tangent cone at $\mathbf{x}^*$ for a convex set S: $T_S(\mathbf{x}^*) = \{\mathbf{y} \mid \exists t > 0 \ni \mathbf{x}^* + t\mathbf{y} \in S\}$, KKT (Karush-Kuhn-Tucker) conditions are said to be satisfied at $(\mathbf{x}^*, \lambda^*)$ iff: $\lambda^* \geq 0, g_i(\mathbf{x}^*) \leq 0 \;\forall\; i, \lambda_i g_i(\mathbf{x}^*) = 0 \;\forall\; i, \nabla f(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla g_i(\mathbf{x}^*) = 0$

Results: The following key theorem:

Theorem 12.2.1. *Let $\boldsymbol{P}$ be a given convex program of the form:*

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

Let $\mathbf{x}^* \in int(X)$ and $f, g_1, \ldots, g_m$ be differentiable at $\mathbf{x}^*$, then:

1. *If the non-linear inequality constraints satisfy the Slater's condition, then $\mathbf{x}^*$ is an optimal solution of $\boldsymbol{P} \Rightarrow$ there exists $\lambda^* \geq 0$ such that the KKT conditions are satisfied at $(\mathbf{x}^*, \lambda^*)$ (and infact, λ^* is an optimal solution of $\boldsymbol{D}$)*

2. *If KKT conditions are satisfied at $(\mathbf{x}^*, \lambda^*)$, then $\mathbf{x}^*$ will be an optimal solution of $\boldsymbol{P}$ (and λ^* will be an optimal solution of $\boldsymbol{D}$ and infact, strong duality will hold).*

(In short, KKT conditions are sufficient for optimality for CPs, and necessary for regular (linear inequalities can be exempted from Slater's condition though) CPs.

12.3 Summary

After reviewing some results from previous lecture, we noted an example of a non-convex $\mathbf{P}$ for which strong duality holds. The problem was that of minimizing (an indefinite i.e., not necessarily convex) homogeneous quadratic over the unit sphere at origin[1]. The Lagrangian function was un-usual (in particular did not satisfy condition of Sion-Kakutani theorem); however had a saddle point — which essentially is the reason for solvability and strong duality.

We then turned our attention to coming up with optimality conditions for constrained optimization problems. Based on the results for unconstrained problems, the following theorem was immediate:

Theorem 12.3.1. *Let **P1** be a program of the form:* $\min_{\mathbf{x}\in X} f(\mathbf{x})$, $\mathbf{x}^* \in int(X)$ *and f is differentiable at $\mathbf{x}^*$. If $\mathbf{x}^*$ is an optimal solution then $\nabla f(\mathbf{x}^*) = 0$. Further in case **P1** is convex, then $\nabla f(\mathbf{x}^*) = 0 \Rightarrow \mathbf{x}^*$ is an optimal solution.*

Once this theorem is in place, proving the key theorem 12.2.1 was not difficult. The proof simply translates the characterization of optimality conditions using saddle point of the Lagrangian into the KKT conditions.

Now using some examples we noted why the case $\mathbf{x}^* \notin int(X)$ is interesting. In particular we noted that gradient at optimal solution being zero is NOT a necessary condition. We argued that the necessary condition in this case is "in the neighbourhood of $\mathbf{x}^*$ which is feasible, the function must not decrease". This lead to the following theorem (which will be proved in the next lecture):

Theorem 12.3.2. *Let **P1** be a convex program of the form:* $\min_{\mathbf{x}\in X} f(\mathbf{x})$, *$\mathbf{x}^* \in X$ and f is differentiable at $\mathbf{x}^*$. $\mathbf{x}^*$ is an optimal solution if and only if $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0 \ \forall \ \mathbf{y} \in T_X(\mathbf{x}^*)$.*

Here $T_X(\mathbf{x}^*)$ is the tangent cone of X at $\mathbf{x}^*$ (we also noted that for convex sets, the tangent cone at any point will be a cone).

[1] Infact, strong duality holds for any minimization of (non-homogeneous) quadratic over a single (non-homogeneous) quadratic constraint

12.4 Recommended Reading

- Section D.2.3 in Nemirovski [2005].
- Section 5.5 in Boyd and Vandenberghe [2004].

Lecture 13

13.1 Key Topics

Proof of generic optimality conditions (non-differentiable, non-interior point), Sufficient conditions for uniqueness of optimal solution, Fenchel duality

13.2 Key Defns. & Results

Definitions: Dual cone; Normal cone; Fenchel dual/Legendre transformation/conjugate of a function;

Results: The following key theorems:

Theorem 13.2.1. *Let **P1** be a convex program of the form:* $\min_{\mathbf{x}\in X} f(\mathbf{x})$ *and* $\mathbf{x}^* \in X$. $\mathbf{x}^*$ *is an optimal solution if and only if there exists atleast one sub-gradient* $(\nabla f(\mathbf{x}^*))$ *at* $\mathbf{x}^*$ *such that* $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0 \;\forall\; \mathbf{y} \in T_X(\mathbf{x}^*)$.

Theorem 13.2.2. *Consider a twice differentiable function* f. *The following statement is true:* $\nabla^2 f(\mathbf{x}) \succ 0 \;\forall\; \mathbf{x}$ *(i.e., Hessian is always pd)* $\Rightarrow f(\mathbf{x}) > f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top(\mathbf{x}-\mathbf{x}_0) \;\forall\; \mathbf{x}, \mathbf{x}_0 \Leftrightarrow f$ *is strictly convex* $\Rightarrow$ *set of minimizers of* f *is either empty or unique*

Theorem 13.2.3. *Let* f *be a function. Then:*

1. f^* *(Fenchel dual of* f*) is a closed convex function.*
2. *If* f *is convex, then* $(f^*)^* = cl(f)$
3. $f^*(0) = -\inf_{\mathbf{x}} f(\mathbf{x})$
4. *If* f *is closed convex, each sub-gradient of* f^* *at origin is a minimizer of* f

13.3 Summary

We began by proving theorem 12.3.2 and later noting theorem 13.2.1. Proving sufficiency condition was easy in both cases and was immediate through first order characterization of convex functions. Necessary condition proof in case of 12.3.2 was by contradiction and that in case of 13.2.1 involves sub-gradient calculus. Proves for theorem 13.2.2 are straight-forward.

We then shifted our attention to duality. For linear programs, we have studied two ways for computing duals: one using LP-Duality, in which case dual can just be written down by inspection; another using Lagrange duality, in which case writing dual involves few steps (rather than simply by inspection). It is interesting that both ways lead to exactly same form of dual for LPs. In optimization theory, researchers have come-up with various notions of duals, which work in various scenarios[1]. In case of common scenarios where there is strong duality with multiple methods, the methods only differ in ease of writing them down and else they are all equivalent. In case of LPs, surely LP duality is easier to write than Lagrange duality. We will study in particular conic duality which is useful for conic programs (we will define later) and touch-upon Fenchel duality: After motivating the definition of Fenchel dual of a function and looking at some examples, we proved theorem 13.2.3. We showed that the usual Lagrangian dual of the problem $\min_{\mathbf{x}} f(\mathbf{x}) s.t. \mathbf{A}\mathbf{x} \leq \mathbf{b}$ can be written using Fenchel's dual of f: $\max_{\lambda \geq 0} f^*(-\mathbf{A}^\top \lambda) - \mathbf{b}^\top \lambda$ and is hence useful in representing Lagrangian duals in a concise manner. More details about Fenchel duality are in Bertsekas [2003] referred below.

13.4 Recommended Reading

- For optimality conditions: prop. 4.6.3, 4.7.1, 4.7.2 in Bertsekas [2003]
- Fenchel duality: sections 3.3, 5.1.6 in Boyd and Vandenberghe [2004]; section C.6.3 in Nemirovski [2005]; chapter 7 in Bertsekas [2003]

[1] Here is a list of named duals: `http://glossary.computing.society.informs.org/second.php?page=duals.html`

Lecture 14

14.1 Key Topics

Conic duality

14.2 Key Defns.

generalized inequalities, pointed cone, QP, QCQP, SOCP

14.3 Summary

The objective was to come up with non-linear generalizations of LPs: $\max_{\mathbf{x}} \mathbf{c}^\top \mathbf{x}$ s.t.$\mathbf{A}\mathbf{x} \leq \mathbf{b}$ such that the LP-duality trick still works. One way of generalizing LPs is by making the linear objective function into a non-linear one and the affine inequalities into non-linear inequalities: $\max_{\mathbf{x}} f(\mathbf{x})$ s.t.$g_i(\mathbf{x}) \leq 0$ (where f, g_i are no more affine). However for this generalization, extension of LP duality lead to Lagrangian duality, which as we noted in previous lecture, is not as easy to write down as an LP dual.

There happens to be another way of generalizing LPs, by generalizing the notion of $\leq$ appearing in the LP and keeping everything else the same. We will be lucky if the generalization of $\leq$ is done in such a way that i) resulting programs have non-affine feasibility sets ii) the simple way of writing down the LP dual still holds. Interestingly, there is such a way of generalizing LP-duality which is called conic duality.

The idea is to pick rules which the $\leq$ satisfies and are used in bounding the primal objective by the dual objective (weak duality): i) $\leq$ is a partial order ii) $\leq$

is preserved under conic combinations. The following theorem was then noted:

Theorem 14.3.1. *A generalized inequality $\preceq$ which defines $\mathbf{v} \preceq \mathbf{u} \Leftrightarrow \mathbf{u}-\mathbf{v} \in K$ where $K = \{\mathbf{x} \mid \mathbf{x} \succeq \mathbf{0}\}$ satisfies the above two conditions (partial order, consistency with conic combinations) if and only if K is a pointed cone.*

Note that if K is chosen as the first quadrant, then the corresponding generalized inequality denoted by $\preceq_K$ is same as the usual $\leq$. If K is chosen as the std. normed cone, then $\begin{bmatrix} \mathbf{x}_1 \\ y_1 \end{bmatrix} \preceq_K \begin{bmatrix} \mathbf{x}_2 \\ y_2 \end{bmatrix} \Leftrightarrow \|\mathbf{x}_2 - \mathbf{x}_1\| \leq y_2 - y_1$. Hence there exist choices of K (pointed cones) such that the feasibility set in the LP above can be made non-linear!

We then noted that the following primal-dual pair will always satisfy weak-duality:

$$\mathbf{P} \quad \max_{\mathbf{x}} \quad \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \preceq_K \mathbf{b}$$

$$\mathbf{D} \quad \min_{\mathbf{y}} \quad \mathbf{b}^\top \mathbf{y} \quad \text{s.t.} \quad \mathbf{A}^\top \mathbf{y} = \mathbf{c}, \ \mathbf{y} \succeq_{K^*} 0$$

where K^* is the dual cone of K. The key advantage being that writing the dual is as simple as that in case of LP, but is not as restrictive as LPs. We noted then that K is a closed pointed cone with non-empty interior if and only if K^* is so. This basically yields symmetry i.e., dual of dual will be primal and strong duality holds whenever K is closed pointed cone with non-empty interior and there exists $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} \prec_K \mathbf{b}$ (regularity condition). These results will be proved in detail in the next lecture.

We then defined QPs, QCQPs, SOCPs and noted that several convex optimization problems (including QPs, QCQPs) can be written as SOCP. Later we will show that duals of SOCPs can be written down using the notion of conic duality.

14.4 Recommended Reading

- Sections 1.3-1.6 in Nemirovski [2005].
- Sections 2.4.1,5.9 in Boyd and Vandenberghe [2004]
- Lobo et al. [1998] is an excellent reference for problems which can be posed as SOCPs.

Lecture 15

15.1 Key Topics

Conic duality

15.2 Key Results & Defns.

Definitions: proper cone, conic program

Results: The following key theorem:

Theorem 15.2.1. *Consider the following primal problem:*

$$\mathbf{P} \quad \max_{\mathbf{x}} \quad \mathbf{c}^\top \mathbf{x} \quad \text{s.t.} \quad \mathbf{b} - \mathbf{A}\mathbf{x} \in K$$

where K is some non-empty set. Consider the following dual problem:

$$\mathbf{D} \quad \min_{\mathbf{y}} \quad \mathbf{b}^\top \mathbf{y} \quad \text{s.t.} \quad \mathbf{A}^\top \mathbf{y} = \mathbf{c},\ \mathbf{y} \in K^*$$

where K^* is the dual cone of K. The following results hold:

1. *Weak duality always holds:* $\mathbf{P} \leq \mathbf{D}$
2. *If K is closed cone with non-empty interior, then dual of dual will be primal (symmetry).*
3. *If K is closed pointed cone with non-empty interior (i.e., K is a proper cone)*[1]*, primal is bounded and primal is strictly feasible (i.e., $\exists \bar{\mathbf{x}} \ni \mathbf{b} -$*

[1] If K is a proper cone then so is K^*. If K is proper cone, then $\mathbf{P}$ is called as a conic program

$\mathbf{A}\bar{\mathbf{x}} \in int(K)$*), then strong duality holds and infact* $\boldsymbol{D}$ *is solvable. By primal-dual symmetry similar statement holds for the dual too.*

4. *If* K *is a proper cone and atleast one of* $\boldsymbol{P}$ *or* $\boldsymbol{D}$ *is bounded and strictly feasible, then* $(\mathbf{x}^*, \mathbf{y}^*)$ *is a pair of optimal primal-dual solutions if and only if:* $\mathbf{c}^\top \mathbf{x}^* = \mathbf{b}^\top \mathbf{y}^*$ *(zero-duality gap) and if and only if:* $(\mathbf{y}^*)^\top (\mathbf{b} - \mathbf{A}\mathbf{x}^*) = 0$ *(complementary-slackness).*

15.3 Summary

The entire lecture was devoted to proving the above "Conic duality" theorem (point no. 4 was left as exercise). Point no. 1 is simple to prove and was done in the previous lecture. For proving point 2, we began re-writing the dual in the form of the primal, while indicating what assumptions on K are sufficient for every step. A crucial result used was: If $K_1, \ldots, K_n$ are cones, then $K = K_1 \times \ldots K_n$ (direct product of cones) is a cone; moreover, $K^* = K_1^* \times \ldots K_n^*$. Proof of point 3 was based on separation theorem.

15.4 Recommended Reading

- Section 1.7 in Nemirovski [2005].

Lecture 16

16.1 Key Topics

Conic duality — LPs, SOCPs, SDPs

16.2 Key Defns.

Semi-Definite Program (SDP), Linear Matrix Inequality (LMI)

16.3 Summary

Lecture began with discussion about three specific choices of K in conic programs, all of which are proper cones and more importantly self-dual: i) $K_l^n = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} \geq 0\}$ (this leads to linear programs) ii) $K_c^n = \{\begin{bmatrix} \mathbf{u} \\ v \end{bmatrix} \mid \|\mathbf{u}\|_2 \leq v, \mathbf{u} \in \mathbb{R}^{n-1}, v \in \mathbb{R}\}$ iii) $K_s^n = \{\mathbf{x} \in S^n \mid \mathbf{x} \succeq 0\}$ (here S^n is the set of all symmetric matrices of size n). It was easy to see the std. LP can be posed as a conic program with $K = K_l^n$. It was interesting to note that an SOCP can be posed as a conic program with $K = K_c^{n_1} \times \ldots K_c^{n_m}$ (and is also self-dual). In particular we noted the dual of an SOCP with two SOC constraints. The natural primal resulting out of the choice $K = K_s$ was called as a Semi-definite program. We in detail motivated the exact form of its primal and dual. The main issue was to generalize the notion of affine transformations for the case $\mathbb{R}^n \mapsto S^m$ and then to generalize the notion of its transpose (adjoint operation, which essentially is

$S^m \mapsto \mathbb{R}^n$). In the following we note the primal and dual form of an SDP:

$$\begin{array}{lll} \mathbf{P} & \max\limits_{\mathbf{x}\in\mathbb{R}^n} & \mathbf{c}^\top \mathbf{x} \\ & \text{s.t.} & \mathbf{B} - \sum_{i=1}^n x_i \mathbf{A}_i \succeq 0 \end{array}$$

where $\mathbf{B}, \mathbf{A}_i \in S^m$. The conic dual of $\mathbf{P}$ is:

$$\begin{array}{lll} \mathbf{D} & \min\limits_{\mathbf{y}\in S^m} & trace(\mathbf{B}\mathbf{y}) \\ & \text{s.t.} & trace(\mathbf{A}_i\mathbf{y}) = c_i,\ \mathbf{y} \succeq 0 \end{array}$$

We then started again with some example problems and stressed on the need to pose various problems in standard forms. We saw examples of problems which can be posed as SOCPs, SDPs (both in primal & dual form). In particular, we can pose any SOCP problem as an SDP. So all the named optimization problems we studied including LPs, QPs, QCQPs, SOCPs can be posed as SDPs and hence are quite generic.

We noted some solvers: i) `SeDuMi`[1]: generic conic program solver which handles all the above three cones. It is open-source. ii) `Mosek`[2]: generic conic program solver which handles all the above three cones. It is proprietary; however student free licence option is available. It is in general far more efficient than `SeDuMi`. In case of both methods, the problems need to be given in standardized LP/SOCP/SDP form iii) `cvx`[3]: optimization problem which are not in standard form also can be given. Very useful for tool especially for initial validation of models. May not be computationally efficient always.

16.4 Recommended Reading

- Sections 2.1, 3.1 in Nemirovski [2005].
- Vandenberghe and Boyd [1996] is an excellent reference on applications of SDP.

[1] Available at http://sedumi.ie.lehigh.edu/
[2] Available at http://www.mosek.com/
[3] Available at http://cvxr.com/cvx/

Lecture 17

17.1 Key Topics

Numerical methods, Gradient-descent

17.2 Key Defns.

Black-box (oracle) setting; zero, first, second order methods; smooth (continuously differentiable) functions; gradient-descent method

17.3 Summary

This was a very general introductory lecture on numerical methods. We discussed terms mentioned above and gave a generic scheme for numerical methods:

1. Pick x^0 (initial guess for optimal solution)
2. At every iteration k, update the guess x^k using information about f at $x^0, \ldots, x^{k-1}$.
3. Repeat until some termination condition (stopping criteria) is met.

Based on what kind of information about f is used, numerical methods are classified into: zero, first, second order methods. We argued that, in general, computing exact solution may not be possible and what we typically look for is an approximate solution. Stopping criteria basically checks if the current solution is good or not. We then discussed various criteria based on: i) $||\nabla f(x^k)|| \leq \epsilon$ (works only for unconstrained problems) ii) duality gap $\leq \epsilon$ iii) KKT gap $\leq \epsilon$ iv) $||x^k - x^*|| \leq \epsilon$ (distance from an optimal solution) v) $dist(x^k, S) \leq \epsilon$ (where S is set of all optimal solutions) vi) $f(x^k) - f(x^*) \leq \epsilon$ (the last three cant be employed in practice, but useful from theoretical view).

We noted the gradient descent scheme (for unconstrained problems):

1. Pick x^0 (initial guess for optimal solution)
2. At every iteration k, update the iterate x^k as $x^{k-1} - s^{k-1}\nabla f(x^{k-1})$.
3. Repeat until some termination condition (stopping criteria) is met.

We gave two different intuitions for this scheme: i) Step in direction of local descent (opp. gradient). s^k has interpretation of step size ii) We are locally approximating the function linearly and minimizing the linear approx while being close to the current iterate i.e., $x^k = \text{argmin}_x x^\top \nabla f(x^{k-1}) + \frac{1}{2s^{k-1}}||x - x^{k-1}||_2^2$ will recover the above update rule. We commented that the scheme can be generalized with various schemes for choosing step sizes s and various descent directions other than $-\nabla f(x^{k-1})$ — each having its own proof of convergence and rate of convergence. In the next lecture we will discuss a particular setting where we show convergence.

17.4 Recommended Reading

- Section 1.1 in Nesterov [2004]

Lecture 18

18.1 Key Topics

Convergence of gradient descent on unconstrained problems with objective in $\mathcal{F}_L^{1,1}$

18.2 Key Defns. & results

Definitions: Notation of $\mathcal{F}_L^{p,q}, p \geq q$.

Results: Theorems 2.1.5, 2.1.6, 2.1.14 in Nesterov [2004]

18.3 Summary

We refreshed notion of Lipshcitz continous functions with examples. Introduced notation $\mathcal{F}_L^{p,q}, p \geq q$ and focussed on the class $\mathcal{F}_L^{1,1}$ which is the set of all smooth convex function whose gradient is Lipshcitz conts. with constant L. We then noted theorem 2.1.6 and subsequently theorem 2.1.5. The former theorem provides a second order characterization of functions of our interest. The latter theorem (in addition to further characterizations), provides important upper and lower bounds on two important quantities: i) error of local linear approximation $(f(y) - f(x) - \nabla f(x)^\top (y-x))$ ii) the inner product $(\nabla f(y) - \nabla f(x))^\top (y-x)$. These form the core of the key theorem in 2.1.14. We then formally proved this key theorem which in addition to proof of convergence, also provided us with rate of convergence: $k \propto O(1/\epsilon)$.

18.4 Recommended Reading

- Sections 2.1.1, thm 2.1.14 (and col 2.1.2) in Nesterov [2004]

Lecture 19

19.1 Key Topics

Convergence of gradient descent on unconstrained problems with objective in $\mathcal{S}^{1,1}_{L,\mu}$

19.2 Key Defns. & results

Definitions: Notation of $\mathcal{S}^{p,q}_{L,\mu}, p \geq q, L \geq \mu$, definition of strongly convex function (zero-order, first order), condition number

Results: Theorems 2.1.11, 2.1.9, 2.1.15 in Nesterov [2004]

19.3 Summary

We introduced the notion of strongly convex functions motivating by the wish to lower bound the error in local linear approx. by a simple quadratic (as gradient's Lip. cont. gave such upper bound). We then provided second-order (theorem 2.1.11) and zero-order (theorem 2.1.9) characterizations of such functions. Discussed few insightful examples. We defined and realized the importance of the condition number $L/\mu \geq 1$. We guessed that lower this value, tighter the bounds on local linear approx., and hence faster the convergence. We proved the condition on the familiar inner-product in thm 2.1.9 and then proved the key theorem 2.1.15. In addition to proof of convergence it provided with rate of convergence: $k \propto O(log(1/\epsilon))$ which is quite faster than rate of convergence in the previous case. The question of whether these rates are "optimal" will be discussed in the next lecture.

19.4 Recommended Reading

- Sections 2.1.3, thm 2.1.15 in Nesterov [2004]

Lecture 20

20.1 Key Topics

Notion of optimal method, example of optimal method

20.2 Summary

For this lecture the reference for theorems/lemmas is always Nesterov [2004].

In the past couple of lectures we arrived at convergence rate of the gradient descent method on two classes of problems: $\mathcal{F}_L^{1,1}$ and $\mathcal{S}_{L,\mu}^{1,1}$. The key question which bothered us was whether this is the optimal convergent rate for the respective classes of problems ? We noted few theorems which answered this question in a convincing fashion.

The first kind of theorems talk about information-based complexity of a class of algorithms i.e., theorems which talk about minimum iterations needed for reaching approximate solution wrt. a particular class of problems and algorithms (refer section 2.1.2 for $\mathcal{F}_L^{1,1}$ and 2.1.4 for $\mathcal{S}_{L,\mu}^{1,1}$). From theorems 2.1.7(2.1.13) it is clear that a minimum of $O(1/\sqrt{\epsilon})$ $(O(\log_{\frac{\sqrt{Q}+1}{\sqrt{Q}-1}}(1/\epsilon)))$ iterations is required for convergence to an ϵ-approximate solution. Though these bounds are far lower than those achieved by gradient descent, it does not still imply that the gradient descent is not optimal.

Nesterov, cleverly constructs an algorithm (which is not a gradient descent or related algorithm) which indeed achieves approximate solution with the above mentioned minimum number of iterations! This algorithm (and variants) are presented in (2.2.6, 2.2.8, 2.2.9 and 2.2.11; pages 76-81). In the lecture we gave a rough intuition and skipped details of convergence etc. (refer theorem 2.2.2).

In particular, this theorem showed that the gradient descent algorithm is not the optimum method.

Lecture 21

21.1 Key Topics

Projected gradient descent, proof of convergence, rate of convergence

21.2 Summary

We discussed an extention of gradient descent algorithm for solving constrained optimization problems of the form $\min_{\mathrm{x}\in X} f(\mathrm{x})$. We assumed $f \in \mathcal{S}^{1,1}_{\mu,L}$ and X is closed convex. As in case of gradient descent, the idea was to minimize a local affine approximation of the function was staying close to the current iterate — ofcourse further constraining that the feasibility set is X. Interestingly, the solution to this problem is same as first taking a usual gradient step and then projecting onto the set X. Hence known as projected gradient. The update can also be interpreted as a descent step along a (non-gradient) direction g^k. We noted that g^k, also called as the gradient mapping, satisfies a cruicial property that gradient also satisfies (eqn. 2.2.7 in Nesterov [2004]) and hence converges with the same rate as in case of an unconstrained problem. Even in this case optimal algorithms can be devised, the details of which we skip. The cruicial point that needs to be remembered is each iteration involves a projection step — which might not be easy for generic convex sets. Hence this method is applicable only for few "simple" feasibility sets.

21.3 Further Reading

1. Sections 2.2.3, 2.2.4 from Nesterov [2004]

Lecture 22

22.1 Key Topics

Non-smooth problems, sub-gradient descent, projected sub-gradient descent, proof and rate of convergence, optimality etc.

22.2 Summary

We considered the case of non-smooth problems i.e., the objective is no longer differentiable. However we do assume its closed convex and hence sub-gradient exists at all relint points of feasibility set. We know that convex functions are locally Lipschitz; moreover we can prove a stronger result — theorem C.4.1 in Nemirovski [2005] — in particular, in any ball in relint of feasibility set, they are Lipschitz conts. We first focus on unconstrained problems.

An obvious strategy is to perform sub-gradient descent. However there are fundamental differences from gradient descent: i) travelling in opposite sub-gradient direction may not (even locally) decrease the function; infact function value might increase! ii) at optimality one cannot insist on sub-gradient being zero. Infact, there is not much information contained in length of sub-gradient (as it can be any value $\leq L$, the local Lipschitz const.). However sub-gradient still satisifies a fundamental property $\nabla f(\mathrm{x})(\mathrm{x}-\mathrm{x}^*) \geq 0$ and hence convergence can be expected. Because of the above mentioned reasons convergences with sub-gradient based methods are expected to be slow and is confirmed by theorem 3.2.1 (Nesterov [2004]) — best possible by such methods is $O(1/\epsilon^2)$! Using results in lemma 3.2.1, we proved thoerem 3.2.2 which establishes the convergence of sub-gradient descent and that it achieves the optimal convergence rate. We extended the results to constrained opt. case (proj. grd. des.) using lemma 3.1.5.

We then gave an overview of kelley's method, bundle methods (proximal, trug-region, level-based) and Newton's method.

22.3 Further Reading

1. Theorem 3.2.1, 3.2.2; Sections 3.2.2, 3.2.3; Lemmas 3.1.5, 3.2.1 from Nesterov [2004]

Lecture 23

23.1 Key Topics

Overview of Interior point algorithms (Barrier method, primal-dual int.pt.) and concept of convex relaxation

23.2 Summary

We gave an overview of interior point algorithms. These are aimed at solving convex problems of the form $\min_{\mathbf{x}} f(\mathbf{x})$ s.t.$f_i(\mathbf{x}) \leq 0$. Algorithms like proj.grad.desc. involve projection step at each iteration which might be difficult in case of large number of non-linear f_i — hence are not suitable. The key idea is to mimic the way constraints are brought to objective in Lagrange theory using a barrier/penalty function. This barrier function acts as a big wall resisting motion from interior of feasibility set to the exterior. We took a logarithmic barrier function and approximated the original problem as an unconstrained problem. We noted that the approximation indeed provides an arbitrarily close approximation to the original problem. However solving the unconstrained problems is difficult especially when good accuracies are desired. To circumvent this, the problem is posed as series of unconstrained problems with improving quality of the barrier function. This so called Barrier method can be proved to be converging and free of numerical problems even when good accuracies are desired. We also noted an primal-dual interior pt. algorithm which updates both the primal and dual variables simultaneously and also uses the same barrier based approximations. In practice primal-dual interior pt. alg. outperform simple int.pt. methods for all standard set-ups.

We then had a brief discussion about convex relaxations. We noted a standard relaxation technique for QCQPs (and hence for many combinatorial opt. problems) and some related results regarding quality of approximation.

23.3 Further Reading

1. Int.pt.alg.: Sections 11.1-11.3, 11.7 in Boyd and Vandenberghe [2004]
2. Conv.Relx.: Sections 3.4.1, 3.5 in Nemirovski [2005]

Bibliography

D. P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, 2003.

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

M. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of Second-Order Cone Programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

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

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

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

L. Vandenberghe and S. Boyd. Semidefinite Programming. *SIAM Review*, 38(1): 49–95, 1996.