Lecture 2: Convex Sets and Their Properties

Rajat Mittal ⋆

IIT Kanpur

The standard form of linear programming is given by,

$$\begin{aligned} \min \quad & c^T x \\ \text{subject to } & Ax = b \\ & x \geq 0 \end{aligned}$$

Using linear algebra, we know the solutions of system of equations, $Ax = b$. In this lecture, Let's take the next step and study set of x's satisfying $Ax = b$ and $x \geq 0$. In other words, what properties does the feasible region of an LP have?

The train of thought starts with the observation that if x_1 and x_2 are two feasible solutions of an LP then $\alpha x_1 + \beta x_2$ is also a solution (given α, β are positive and add up to 1). These are called the convex combinations of x_1 and x_2.

The study of these feasible regions will take us through linear combinations (convex combinations are special cases), sets arising from such combinations and their properties.

1 Linear combinations

We denote the set of real numbers as $\mathbb{R}$. We will mostly work with the vector space $\mathbb{R}^n$; its elements will be called vectors. Remember that a vector space is a set of vectors closed under addition and scalar multiplication.

While studying linear algebra, the set of equations Ax were viewed as a linear combination of columns of x. Actually, there are multiple ways to combine a given set of vectors. Let us look at a few important ones.

For vectors $x_1, x_2, \cdots, x_k$, any point y is a linear combination of them iff

$$y = \alpha_1 x_1 + \alpha_2 x_2 \cdots + \alpha_k x_k \quad \forall i,\ \alpha_i \in \mathbb{R}.$$

Exercise 1. What do you get by taking linear combination of three or in general n points?

Hint: Will you get a full dimensional vector space?

Restricting α_i's to be positive, we get a conic combination,

$$y = \alpha_1 x_1 + \alpha_2 x_2 \cdots + \alpha_k x_k \quad \forall i,\ \alpha_i \geq 0 \in \mathbb{R}.$$

Instead if we put the restriction that α_i's sum up to 1, it is called an affine combination,

$$y = \alpha_1 x_1 + \alpha_2 x_2 \cdots + \alpha_k x_k \quad \forall i,\ \alpha_i \in \mathbb{R},\ \sum_i \alpha_i = 1.$$

When a combination is affine as well as conic, it is called a convex combination,

$$y = \alpha_1 x_1 + \alpha_2 x_2 \cdots + \alpha_k x_k \quad \forall i,\ \alpha_i \geq 0 \in \mathbb{R},\ \sum_i \alpha_i = 1.$$

Exercise 2. What is the geometric shape obtained by taking linear/conic/affine/convex combination of two points in $\mathbb{R}^2$?

⋆ The contents of this lecture note are inspired by the books from Boyd and Vandenberghe, Dantzig and Thapa, Papadimitriou and Steiglitz.

We notice that only linear combination gives us a subspace. An affine combination gives a line, a conic combination gives us a line segment and a conic combination gives a cone-shape in two dimensional space.

So, these combinations give rise to *special* subsets/subspaces of the original vector space. You will see that such sets naturally arise while studying feasible regions of a linear program.

1.1 Affine sets

An affine combination of two points, in two dimension, gave a *line*. The following definition generalizes line to higher dimension.

Definition 1. *Affine set: A set S is called affine iff for any two points in the set S, the line through them is contained in S. In other words, for any two points in S, their affine combination is in the set S too.*

Exercise 3. Show that line is an affine set.

Theorem 1. *A set S is affine iff any affine combination of points in the set (countable points) is also in S.*

Proof. Exercise (use induction). □

Exercise 4. What is the affine combination of three points?

We will try to answer this question methodically. Suppose, the three given points are x_1,x_2 and x_3. Then, any affine combination of these three points can be written as

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3, \ \sum_i \alpha_i = 1.$$

We can simplify the expression to,

$$\alpha_1(x_1 - x_3) + \alpha_2(x_2 - x_3) + x_3.$$

Notice that there are no constraints on α_1, α_2. We can think of this sum as,

$$x_3 + (\text{plane/space generated by } (x_1 - x_3) \text{ and } (x_2 - x_3)) .$$

In other words, it is a linear subspace shifted by x_3.

Exercise 5. What is the dimension of this linear subspace?

Generalizing the previous argument, any affine set S can thought of as an offset x added to some vector space V. Hence,

$$S = \{v + x : \ v \in V\}.$$

You can also prove that such a set is affine (exercise). Not just that, given an affine set S and a point x inside it, the set

$$V = \{s - x : \ s \in S\}$$

is a vector space. In the assignment, you will show that the vector space associated with an affine set does not depend upon the point chosen. So, we can define the dimension of the affine set as the dimension of the subspace.

Exercise 6. Show that the feasible region of a set of linear equations is affine.

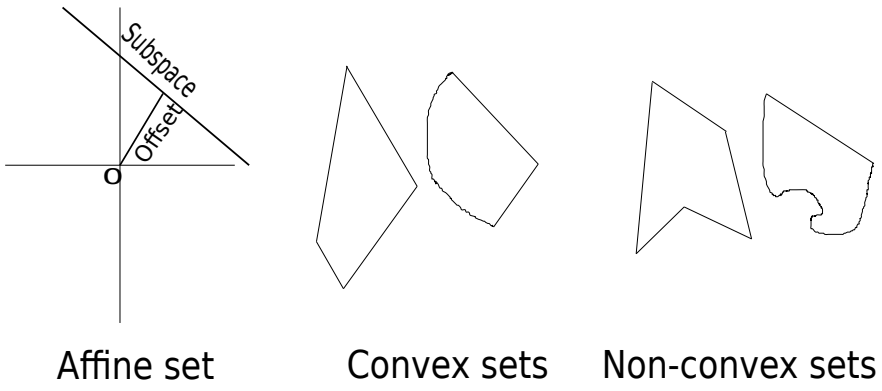

Fig. 1. Different kind of sets

1.2 Convex set

From the definition of affine sets, we can guess the definition of convex sets.

Definition 2. *A set is called convex iff any convex combination of a subset is also contained in the set itself.*

Theorem 2. *A set is convex iff for any two points in the set their convex combination (line segment) is contained in the set.*

We can prove this using induction. It is left as an exercise.

Convex hull: Convex hull of a set of points S, denoted $Conv(S)$, is the set of all possible convex combinations of the subsets of S. From the definition, it is clear that the convex hull is a convex set.

Theorem 3. *$Conv(S)$ is the smallest convex set containing S.*

Proof. Suppose there is a smaller convex set C containing S, then C contains all possible convex combinations of S. Hence, C contains $Conv(S)$. Then C is bigger than $Conv(S)$, contradiction. This implies $C = Conv(S)$. □

Convex hull of S can also be viewed as the intersection of all convex sets containing S (Prove it).

Exercise 7. What is the convex hull of three vertices of a triangle? What if I add a point inside the triangle? What about outside?

Intersections and unions of convex sets Suppose we are given two convex sets S_1 and S_2. What happens when we take their intersection or union?

Intersection of two convex sets is convex too. To prove it, consider two points in the intersection $S_1 \cap S_2$. Since both points are in S_1, S_2, the line segment connecting them is also contained in both S_1, S_2. This proves that the intersection is also convex. The same argument can be extended to the intersection of finite number of sets and even too infinite number of sets.

Though, the same property does not hold true for unions. In general, union of two convex sets is not convex. To obtain convex sets from the union, we can take convex hull of the union.

Exercise 8. Draw two sets such that there union is convex. Draw two convex sets, s.t., there union is not convex. Draw the convex hull of the union.

2 Important affine and convex sets

2.1 Lines

The simplest example of a non-trivial affine set is probably a line in the space $\mathbb{R}^n$. Given two points, x_1 and x_2, a line through x_1 and x_2 is the set of all points y of the form,

$$y = \alpha x_1 + (1-\alpha)x_2, \quad \alpha \in \mathbb{R}.$$

Specifically, this gives us the unique line passing through x_1 and x_2. If we constraint α to be in $[0, 1]$, it is called the line segment between x_1 and x_2.

Exercise 9. Prove that the line segment is a convex set.

Equivalently, a point is on the line segment between x_1 and x_2 iff it is a convex combination of the given two points. Note that the condition for being a convex set is weaker than the condition for being an affine set. Hence an affine set is always convex.

Since line is an affine set, it is a convex set. A line segment will be a convex set but not an affine set.

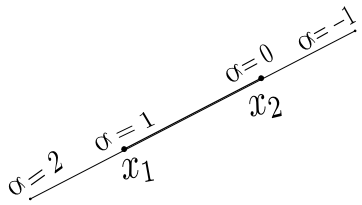

Fig. 2. A line, line segment, points on the line and α's corresponding to them

There is another way to think of a line in $\mathbb{R}^2$. You will prove in the assignment that every point on a line has same inner product with the vector perpendicular to $x_1 - x_2$. In other words, the equation of line can be written as,

$$a^T x - b = 0.$$

In the previous equation, a is the vector perpendicular to $x_1 - x_2$.

Exercise 10. What is b?

2.2 Hyperplane and Halfspaces

We extend the idea of a line to bigger dimensions, this gives rise to the concept of hyperplanes. A hyperplane H is described by a vector $a \in \mathbb{R}^n$ and a number $b \in \mathbb{R}$

$$H = \{x : a^T x - b = 0\}$$

Exercise 11. Prove that a hyperplane is affine and so convex too.

Suppose, we know a point x_0 on the hyperplane H. The previous equation can be changed to

$$a^T x = b \Rightarrow a^T x = a^T x_0, \; x_0 \in H.$$

We can view the hyperplane as the set of all points which have same inner product b with the vector a. This gives a very nice geometrical picture of the hyperplane, i.e., all points in H can be expressed as the sum of x_0 and a vector orthogonal to a (say $a^\perp$). We get another definition of hyperplane as

$$H = \{x_0 + a^\perp : \ a^T a^\perp = 0\}.$$

Notice, this definition assumes that we know a point on the hyperplane. Though, this point is not special in any way; for any point on the hyperplane we can define the hyperplane in a simialr way.

Vector a is called the *normal* vector of the hyperplane and b is called the *offset*. Another way to think of this hyperplane is: take the set of all vectors orthogonal to a (hyperplane, passing through origin) and offset them by distance $\frac{b}{\|a\|}$ (look at Fig. 3).

A hyperplane divides the space into two parts, $a^T x \geq b$ and $a^T x \leq b$. Geometrically, they are the two sides of the hyperplane. Is a halfspace affine/convex?

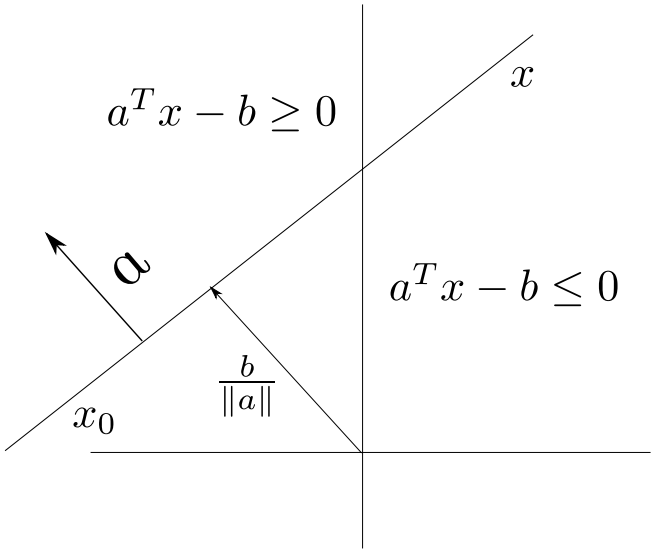

Fig. 3. Halfspace with normal vector a and offset b .

2.3 Polytopes and Polygons

Looking at the format of a linear program, we will mostly be interested in linear equations and linear inequalities (our constraints). Remember that the set of points which satisfy the set of all constraints is called the *feasible region*. When these constraints are linear inequalities and linear equalities, the feasible region is called a polytope.

Mathematically, S is a polytope iff

$$S = \{x : a_i^T x - b_i \leq 0 \ \ i = 1, 2, \cdots, m \text{ and } c_i^T x - d_i = 0 \ \ i = 1, 2, \cdots, p\}$$

A bounded polytope in two dimensions is called a polygon.

Exercise 12. Prove that the polytope is a convex set.

Exercise 13. What are the polytopes in one dimension?

Geometrically, polytopes are intersections of hyperplanes and halfspaces. Since intersection of convex sets is also convex, a polytope is convex.

Remember that a polytope could be bounded or unbounded. Intuitively, polytopes have vertices, edges, planes and hyperplanes as their bounding surface.

Exercise 14. Find a set of equalities and inequalities, s.t., the corresponding polytope is unbounded.

A bounded polytope can be thought of as the convex hull of its vertices. Alternately, a bounded polytope can have an alternate definition as the convex hull of a finite set of points.

The statement that these two definitions (the previous one and the original definition in terms of equalities and inequalities) are the same is known as Minkowski-Weyl theorem. The proof of this theorem is out of the scope of this course.

We haven't defined vertices/extremal points formally till now. It is intuitively clear that a vertex is a corner of the polytope. Formally, A vertex of a polytope is the point which cannot be expressed as the convex combination of two different points in the polytope. This implies if there is a line segment containing the vertex inside the polytope, then the vertex is an end-point of that line segment.

We saw that the convex hull of a triangle and a point inside the triangle is triangle itself. Suppose we are given a convex set and we want to find the minimal set whose convex combinations will generate the entire set.

By the definition of vertices, all vertices should be in this minimal set. It turns out that the for a bounded polytope, this set (set of all vertices) is enough.

Exercise 15. Suppose $S = Conv(x_1, \cdots, x_k)$. Prove that x_i is not extremal/vertex if and only if it can be written as the convex combination of other x_j's.

2.4 Cones

We have seen sets made by vertices, lines, line segments. Now we look at sets generated by *rays*.

Exercise 16. What is a ray?

A set is called a cone iff every ray from origin to any element of the set is contained in the set. Hence, a set C is a cone iff for every $x \in C$ we have $\alpha x \in C, \alpha \geq 0$.

Note 1. A cone is *not* a set which has all possible conic combinations of all its points (show an example). To remind you the notion of conic combination, a conic combination of vectors $x_1, \cdots, x_k \in \mathbb{R}^n$ is any vector of the form $\alpha_1 x_1 + \cdots + \alpha_k x_k$ for $\alpha_1, \cdots, \alpha_k \geq 0$.

The previous paragraph implies that a cone is not necessarily convex (give example of a cone which is not convex). A set which is a cone and is convex is called a convex cone. In this course we will mostly be concerned with convex cones.

Mathematically, a convex cone C is a cone where,

$$\forall x_1, x_2 \in C \text{ and } \alpha_1, \alpha_2 \geq 0,\ \alpha_1 x_1 + \alpha_2 x_2 \in C.$$

So, a cone is convex iff it contains all the conic combinations of its elements.

Clearly the set $Cone(x_1, x_2, \cdots, x_k) := \{\alpha_1 x_1 + \cdots + \alpha_k x_k :\ \forall i\ \alpha_i \geq 0,\ x_i \in \mathbb{R}^n\}$ is a convex cone. It is called a *finitely generated cone* because it is generated by finite number of vectors.

Note 2. Some authors define cones as sets closed under *positive* scalar multiplication. We have defined cones as sets closed under non-negative scalar multiplication.

A convex finitely generated cone is also a polytope, i.e., it can be represented in terms of linear equalities and inequalities. Next theorem gives a way to construct this set of linear equalities and inequalities.

Theorem 4 (Weyl). *A non-empty finitely generated convex cone is a polytope.*

Proof. Suppose the set of generators for cone C are $x_1, \cdots, x_k$. Define a matrix X with x_i as the i-th column; X has k columns. Then cone C can also be described as,

$$C = \{x : x = X\alpha,\ \alpha \in \mathbb{R}^k_+\}.$$

Converting equalities into inequalities

$$C = \{x : x - X\alpha \leq 0,\ X\alpha - x \leq 0,\ -\alpha \leq 0\}$$

Notice that all the inequalities are of similar kind, the constant term is zero. Now, we can eliminate α from these inequalities using something known as *Fourier-Motzkin elimination* (next lemma). As a bonus, inequalities will have the same structure, no constant term.

Lemma 1. *Let $Ax \leq b$ be a system of m inequalities in n variables. This system can be converted into another equivalent system $A'x \leq b'$, with $n-1$ variables and polynomial in m many inequalities. Here equivalent means:*

- *Any solution x of old system will be a solution of the new system ignoring the removed variable.*
- *Given any solution x of new system ($A'x \leq b'$), we can find a solution (x_0, x) of old system.*

Proof. Suppose the variable to be removed is x_0. We divide all the inequalities into three sets depending upon whether the coefficient of x_0 is positive (P), negative (N) or zero (Z). Divide the inequalities in P and N by the modulus of the coefficient of x_0.

The inequalities in the new system are the inequalities from Z and every inequality of the form

$$P_i + N_j \leq 0,\ P_i \in P,\ N_j \in N.$$

Exercise 17. Prove that the construction above satisfies the first condition. Also, convince yourself that if b was zero initially, the final system will not have non-zero constant terms.

For the more complicated case, we need to show that any solution of new inequalities can be extended (with x_0) to a solution of previous inequalities.

Fix a solution of new inequalities, say y. Define $P_i(y) := p_i$ and $N_j(y) := n_j$ for all $i \in P$ and $j \in N$.

We need to find x_0, s.t., $x_0 + p_i \leq 0$ and $-x_0 + n_j \leq 0$ for all $i \in P$ and $j \in N$. That is,

- $x_0 \leq \min_{i \in P} -p_i$.
- $x_0 \geq \max_{j \in N} n_j$.

Such an x_0 always exists because feasibility of y implies

$$\min_{i \in P} -p_i \geq \max_{j \in N} n_j,$$

showing the second condition of the lemma. □

Using the previous lemma multiple times, α can be eliminated from the system of equations defining the cone. As a result, we get

$$C = \{x : Ax \leq 0\},$$

proving that C is a polytope. □

Note 3. A general polytope is $Ax \leq b$ and we will see that a finitely generated cone is $Ax \leq 0$.

2.5 Characterization of polytopes

We saw that cones and bounded polytope have two representations. One in terms of linear inequalities and one in terms of convex combinations (conic combination). There is a similar theorem for polytope.

Theorem 5. *Let there be a polytope defined by a set of inequalities, $P = \{x \in \mathbb{R}^n : Ax \leq b\}$. Then, there exist vectors $x_1, \cdots, x_k \in \mathbb{R}^n$ and $y_1, \cdots, y_l \in \mathbb{R}^n$, s.t.,*

$$P = Cone(x_1, \cdots, x_k) + Conv(y_1, \cdots, y_l)$$

Note 4. The sum defined here is Minkowski's sum. In this case, $S_1 + S_2 = \{x_1 + x_2 : \ x_1 \in S_1, \ x_2 \in S_2\}$

This is known as *affine Minkowski-Weyl* theorem. We will not cover the proof of this theorem in this course.

Notice that there are equivalent characterizations of cone/polytope/bounded polytope in terms of convex/conic hulls or linear inequalities. It is instructive to remember the special forms of linear inequalities and hulls required to make these shapes.

	Linear inequalities	Convexity
Bounded polytope	Bounded and $Ax \leq b$	$Conv(y_1, \cdots, y_n)$
Finitely generated Cone	$Ax \leq 0$	$Cone(y_1, \cdots, y_n)$
Polytope	$Ax \leq b$	$Cone(y_1, \cdots, y_n) + Conv(z_1, \cdots, z_m)$

3 Convex functions

A function is called convex if the line segment connecting any two points on the graph lies above the graph. Formally, a function $f : \mathbb{R}^n \to \mathbb{R}$ is called convex iff

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y), \ \forall x, y \in \mathbb{R}^n, \ 0 < \theta < 1.$$

In this case, domain is assumed to be entire $\mathbb{R}^n$. In general, if the domain is a convex subset of $\mathbb{R}^n$ satisfying the above mentioned property, it is called a convex function. The canonical example of a convex function is $f(x) = x^2$. You should check that this function is convex.

Note that since the domain is convex, if we restrict the function on any line passing through the domain, the restricted function will be convex. Conversely, if the function is convex on all the lines passing through the domain, then it is convex on the whole domain.

A function f is called concave if it satisfies the above mentioned inequality in opposite direction,

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta)f(y), \ \forall x, y \in \mathbb{R}^n, \ 0 < \theta < 1.$$

It is clear that if f is convex then $-f$ is concave.

Exercise 18. Characterize the functions which are both convex and concave.

Relation to convex sets A convex set is the set which contains all possible convex combinations of its points. What is the relation between convex sets and convex functions?

The *epigraph* of the function is the set of all points which lie above the graph of the function. Formally, given a function $f : \mathbb{R}^n \to \mathbb{R}$, the epigraph of f is the set,

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

If the function is convex, then its epigraph is a convex set and vice versa. This gives a geometric interpretation of a function being convex. The set of all points lying below the graph of a function, oppposite side of the epigraph, is known as the *hypograph* of that function.

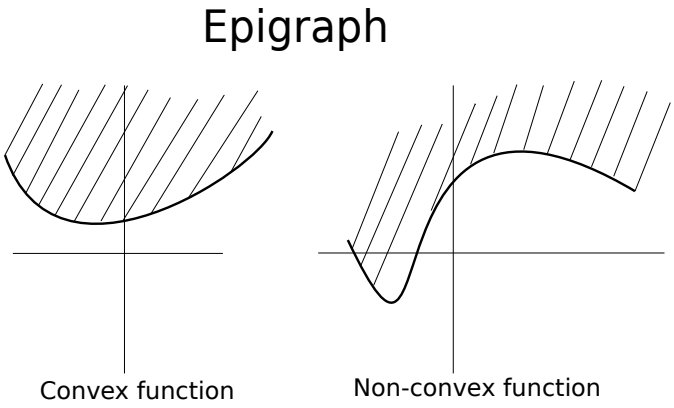

Fig. 4. Epigraph for functions

3.1 Extra reading: Jensen's inequality

One of the most important properties of convex function is *Jensen's inequality*. Given a random variable X and a convex function f,

$$E(f(X)) \geq f(E(X))$$

This can be proved using the straightforward extension of the basic inequality for convex functions.

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

This innocent looking inequality can be used to prove other inequalities by using different convex functions.

- AM-GM: Since the function $\log x$ is concave (it satisfies Jensen's inequality in opposite direction). Applying Jensen's inequality on it for a random variable which takes values a and b with equal probability.

$$\begin{aligned} \log \frac{a+b}{2} &\geq \frac{\log a + \log b}{2} \\ \Rightarrow \quad \frac{a+b}{2} &\geq \sqrt{e^{\log a + \log b}} \\ \Rightarrow \quad \frac{a+b}{2} &\geq \sqrt{ab} \end{aligned}$$

  This shows the arithmetic mean-geometric mean inequality.
- Suppose $x_1, \cdots, x_k$ are some positive numbers. Say $p_1, \cdots, p_k$ is a probability distribution ($\sum_i p_i = 1,\ p_i \geq 0$). Then we need to show,

$$\Pi_i x_i^{p_i} \leq \sum_i p_i x_i.$$

  Proof. Assume that $x_i = e^{y_i}$ (why can we assume that?). Then the inequality turns into,

$$\Pi_i e^{\sum_i y_i p_i} \leq \sum_i p_i e^{y_i}.$$

  This is Jensen's inequality for function $f(x) = e^x$. □

3.2 Optimization with convex functions

Remember that our feasible set is convex (intersection of convex sets). We are optimizing a linear function, which is both convex as well as concave.

It is comparatively easier to optimize a convex function, as compared to a non-convex function, due to the structure present in these functions. There are essentially two properties which are very helpful for optimization.

- A local minimum is a global minimum.
- Maximum is found on the boundaries.

A similar statement can be given for concave function optimization by changing maximum/minimum.

Global optimality and local optimality For a general function there are two kind of optimal points. There is a *global optimum*, which is the general notion of being optimal as compared to any other point on the domain. Hence, a point x^* is globally minimum for a function f iff $f(x) \geq f(x^*)$ for any x in domain of f. In general, it is very hard to find such points.

On the other hand, there is another concept of a *local optimum*, where the point has minimum value among all values in the neighborhood. Hence, a point x^* is a local optimum point for a function f iff for some ϵ,

$$f(x^*) \leq f(x), \ \forall x \in B_\epsilon(x^*).$$

Here, $B_\epsilon(x^*)$ is the ball of radius ϵ around x^*. As compared to global optimum, it is much easier to check for a local optimum. For instance, we can check that the derivative in every direction is positive and continuous.

The special property of convex functions is that a locally minimum point is globally minimum too (feasible region needs to be convex). This makes it much easier to optimize a convex function. The proof is given for a single variable function, but the theorem works for multi-variable function too.

Proof (local minimum is global). We will use another characterization of a function being convex. A function is convex iff for any two points x, y in the domain

$$f(y) \geq f(x) + f'(x)(y - x)$$

This definition holds for functions of one variable. In the case of more variables, $f'(x)$ will be a vector and a dot product with $y - x$ can be taken.

Suppose point x^* is locally optimal. Since the function is convex, for any point y,

$$f(y) \geq f(x^*) + f'(x^*)(y - x^*).$$

This can be written as,

$$f(y) \geq f(x^*) + \left(\lim_{t \to 0} \frac{f(x^* + t(y - x^*)) - f(x^*)}{t(y - x^*)} \right) (y - x^*),$$

using the definition of first derivative. This can be simplified to,

$$f(y) \geq f(x^*) + \left(\lim_{t \to 0} \frac{f(x^* + t(y - x^*)) - f(x^*)}{t} \right).$$

Clearly the second term on the right is positive because x^* is a local minimum. This implies, for any y in the domain,

$$f(y) \geq f(x^*)$$

So x^* is globally optimal too. □

Optimality at the boundary For a convex function it can be shown that the maximum always lies at the extremal points of the feasible region. Similarly for a concave function the minimum is always attained at the extremal points of the feasible region.

Exercise 19. Prove the above statement.

Suppose that the feasible region is bounded and convex. Given that a linear function is both convex and concave,

- the optimal for a linear program lies on its extremal points,
- local optima is a global optima.

From the previous discussion, *vertices* are the extremal points of the feasible region of LP. Hence, the optimal of an LP lies on its vertex.

Note 5. This does not mean all optimal solutions lie on vertex. Just that, there always exist at least one optimal solution on the vertex. In general, an entire edge or a face could give us the optimal solution.

4 Simplex method: informal idea

Let us look at few of the properties discussed in the previous section about linear programs. For the sake of simplicity, we start with the case when the feasible region is bounded.

- The local optimum is the global optimum.
- The feasible region of the LP is a polytope whose extremal points are vertices.
- The optimum of the LP will lie on a vertex.

These properties suggest a pretty simple algorithm. Say, we want to maximize an objective function. Start with a vertex, check if it is locally optimum (just looking at directions corresponding to the edges from the vertex is enough). How do we find a starting vertex? We will answer it later.

If the vertex is not optimal, there is a direction where objective increases, follow that direction. Moving in the direction of the increasing objective function, find a vertex. Keep repeating this process untill you find a local optimum. The algorithm will terminate since the number of vertices are finite and the objective keeps increasing (we can't cycle back to a vertex).

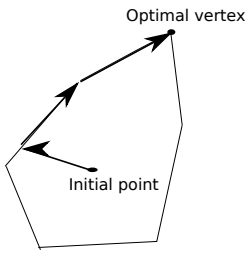

Fig. 5. Outline of simplex method

To clarify the algorithm more, we need to answer a few questions.

- How to find a vertex? What is the mathematical description of a vertex?
- How to check every interesting direction (corresponding to all the edges)?
- How can we find the next vertex?

The *simplex algorithm* answers all these questions and hence solves the LP for us. In general, the number of vertices can be exponential in the size of the input (the number of variables). We will see this later. Since, in worst case, we might have to go through all the vertices, simplex algorithm might not be efficient.

After more than 60 years of research, we still don't know if simplex algorithm is polynomial time. Almost all the known implementations have counterexamples which run for exponential time.

The simplex algorithm was developed by George Dantzig in 1947. In spite of the caveat that it can't be proved efficient theoretically, it is considered to be one of the most important algorithmic achievements in the 20th century. Many industries still prefer to use this method over other known polynomial time solutions.

4.1 Vertex as BFS

How do we know if a particular feasible point is a vertex? Consider the feasible region for the standard form,

$$R = \{x : Ax = b,\ x \geq 0\}$$

We assume that there are n variables and m equality constraints, i.e., A is an $m \times n$ matrix. We can assume that number of constraints (m) is less than n (why?).

To simplify the notation, if $B \subseteq [n]$ and $x \in \mathbb{R}^n$, x_B denotes the restriction of x on the co-ordinates of x. We can similarly define A_B, where we only consider columns of A indexed by elements in B.

A point $p \in \mathbb{R}^n$ is a basic feasible solution (BFS), if its co-ordinates can be divided into two parts (B and N), s.t.,

1. $p_i = 0$ for all $i \in N$.
2. $A_B p_B + A_N p_N = b$ and A_B is invertible. Remember, A_B, p_B and A_N, p_N are the restrictions of A, p to the columns of B and N respectively.

Note 6. We are allowed to permute the columns of A to get such a representation. Since A_B is invertible, p_B is the unique solution of $A_B x = b$

Exercise 20. What is the size of B?

A *basic feasible solution (BFS)* provides the characterization of the vertices.

Theorem 6. *A point $p \in \mathbb{R}^n$ is a vertex iff it is a BFS.*

Proof. ($\Leftarrow$) Suppose p is a BFS but not a vertex. Then there exist q, r, s.t., $Aq = Ar = b, q \geq 0, r \geq 0$ and $\theta q + (1-\theta)r = p$ for some $\theta \in [0, 1]$. Since $q, r, \theta, 1-\theta$ are positive, the co-ordinates corresponding to N for q, r should be zero. This implies there are three different solutions to $A_B x = b$ (namely p_B, q_B, r_B), contradicting the uniqueness of p_B.

($\Rightarrow$) Suppose p is a vertex. Let S denote the set of columns of A for which corresponding entry in p is not zero. We need to show that A_S is invertible. Equivalently, we will show that the columns in S form a linearly independent set. If not, there exist α_i's, not all zero, s.t.,

$$\sum_{i \in S} \alpha_i c_i = 0.$$

Here, c_i is the i^{th} column of A.

Call the vector of these coefficients α_i's as α. Then, $p + t\alpha$ is a feasible solution for $Ax = b$ when absolute value of t is sufficiently small. Notice that $p + t\alpha$ will always satisfy $Ax = b$. Though, for a big t, $p + t\alpha$ might not be non-negative.

Exercise 21. Show that there exist ϵ, s.t., $p + t\alpha$ is feasible if $|t| \leq \epsilon$.

Take t to be in that range. The point p can be written as $p = \frac{1}{2}(p + t\alpha) + \frac{1}{2}(p - t\alpha)$, hence it is not a vertex.

Exercise 22. Show that the number of independent columns in A is m.

From the previous exercise, $|S| \leq m$. If $|S| = m$, then A_S is invertible (full rank) and we are done. If $|S| < m$, then just add any extra columns of A to make S full rank. □

Note 7. Since $|B| = m$, Thm. 6 implies that any BFS will have at most m non-zero co-ordinates.

Given a handle on vertices, we will show how to progress in the simplex method using an example. The strategy will be formalized later.

4.2 A simple example

We will start with a linear program, s.t., it is easy to find one vertex. Take the linear program,

$$\begin{aligned} \max \quad & 2x_1 - x_2 \\ \text{s.t.} \quad & x_1 + x_2 \leq 2 \\ & x_1 - x_2 \leq 1 \\ & x \geq 0. \end{aligned}$$

We introduce *slack variables* to turn inequalities into equalities,

$$\begin{aligned} \max \quad & 2x_1 - x_2 \\ \text{s.t.} \quad & x_1 + x_2 + x_3 = 2 \\ & x_1 - x_2 + x_4 = 1 \\ & x \geq 0. \end{aligned}$$

Luckily b_i's are both positive and hence give us a BFS, $x_1 = 0, x_2 = 0, x_3 = 2, x_4 = 1$. How should we move to another vertex with better objective value?

We will write the constraints by expressing basic variables in terms of non-basic ones.

Exercise 23. Why can we do that?

Using these equations, the objective function can be written *just* in terms of non-basic variables.

$$\begin{aligned} \max \quad & 0 + 2x_1 - x_2 \\ \text{s.t.} \quad & x_3 = 2 - x_1 - x_2 \\ & x_4 = 1 - x_1 + x_2 \\ & x \geq 0. \end{aligned}$$

The objective value for this case is 0 and it is clear that x_1 can be increased to get a better value. Increasing x_1 is like moving on an edge.

Can we increase x_1 arbitrarily? x_1 is bounded by 1 because of the second constraint. So, we remove x_4 from our basic variables (*leaving variable*) and put x_1 in (*entering variable*). The new BFS is $x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 0$. Let us substitute $x_1 = 1 - x_4 + x_2$ to get new form of equations.

$$\begin{aligned} \max \quad & 2 - 2x_4 + x_2 \\ \text{s.t.} \quad & x_3 = 1 - 2x_2 + x_4 \\ & x_1 = 1 - x_4 + x_2 \\ & x \geq 0. \end{aligned}$$

The objective value is 2 for this solution and can be increased by increasing x_2. The variable x_2 can be increased up to only $\frac{1}{2}$ because of the first constraint. The new BFS is $x_1 = \frac{3}{2}, x_2 = \frac{1}{2}, x_3 = 0, x_4 = 0$. Again, using the value of x_2 in terms of x_3, x_4 (first constraint),

$$\begin{aligned} \max \quad & \tfrac{5}{2} - \tfrac{3}{2}x_4 - \tfrac{1}{2}x_3 \\ \text{s.t.} \quad & x_2 = \tfrac{1}{2} - \tfrac{1}{2}x_3 + \tfrac{1}{2}x_4 \\ & x_1 = \tfrac{3}{2} - \tfrac{1}{2}x_4 - \tfrac{1}{2}x_3 \\ & x \geq 0. \end{aligned}$$

It is clear that this is the optimal solution from the form of objective function (all coefficients of objective function are negative). This shows that $x_1 = \frac{3}{2}, x_2 = \frac{1}{2}, x_3 = 0, x_4 = 0$ is optimal (why?).

Note 8. In case the feasible region is not bounded, there is a possibility that travelling in the direction of an extremal ray takes us to an unbounded optimal solution.

5 Simplex method: formal details

Before explaining formally what simplex method is, please note that there are many simplex algorithms. All of them follow the same ideas we discussed in the previous section. The form given below is a particular implementation of it.

5.1 Simplex algorithm assuming a starting vertex

Let us look at the simpler case: a starting vertex is given to us to initiate simplex algorithm. We just need to see how to move from one vertex to another (one BFS to another) and finally check if the vertex is optimal.

Remember that the LP we are trying to solve is,

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0. \end{aligned}$$

Suppose at a particular moment the BFS given to us is, $p = (p_B, p_N), p_N = 0$. Let us look at the feasibility equations,

$$A_B x_B + A_N x_n = b$$

Since A_B is invertible, we multiply this equation with A_B^{-1}.

$$x_B + A_B^{-1} A_N x_n = A_B^{-1} b$$

In other words, all the *basic variables* (variables in set B) can be expressed in terms of *non-basic variables* (variables in set $[n] - B$). Using this, the cost function can also be expressed as the function of only non-basic variables.

Exercise 24. Why can this transformation be useful?

We assume that the BFS, feasibility equations and cost function are given in this form.

- Every basic variable is expressed in terms of non-basic variables.
- The cost vector is just a function of non-basic variables.

From this form, it is clear that the objective value for p is just the constant term in cost function. Also, the current BFS is optimal if the coefficients of all the non-basic variables in the cost function are negative. Hence, we just need to find a way to move from one BFS to another (and change the form of equalities) such that the objective value does not decrease.

Suppose at any stage the set of basic variables is $\mathcal{B}$ and set of non basic variables is $\mathcal{N}$. Say the equations and the objective function looks like,

$$\begin{aligned} \max \quad & C' + \textstyle\sum_{j \in \mathcal{N}} c'_j x_j \\ \text{s.t.} \quad \forall\, i \in \mathcal{B} \quad & x_i = b'_i - \textstyle\sum_{j \in \mathcal{N}} a'_{ij} x_j \\ & x \geq 0. \end{aligned}$$

The *prime* indicates that the value of these coefficients have changed from the starting iterations. How do we move to next iteration?

Exercise 25. How to check if this is the optimal BFS?

If all the c'_j are negative then we have arrived at the optimal solution. Otherwise, choose the variable with the highest c'_j (arbitrary choice, any variable with positive c'_j will work). That variable will move from non-basic to basic (it is called the *entering variable*).

While increasing it gradually, some basic variable will become negative. The first one to do so, say x_i, will be the *leaving variable*. That means, it will move from basic to non-basic. The new values of the basic solution can be calculated easily (How?, Exercise).

In case all the coefficients for the entering variable are positive in the constraints, then the optimal is unbounded and algorithm stops.

Exercise 26. In terms of a', c and b', come up with the definition of entering and leaving variables.

The only task is to write the equations in the required form (in terms of non-basic variables). Take the equation where x_j is expressed as the function of x_i and other non-basic variables. Change it to an equation, where x_i is expressed in terms of x_j and other basic variables. Use this equation to substitute everywhere the value of x_i in terms of x_j. Hence, we will get the equations and the objective function in the required form according to the new BFS.

Exercise 27. Why is the new solution a BFS. In other words, why is new A_B invertible?

There are two things missing in our description of the algorithm.

- How to find the starting vertex?
- What if one of the b'_i is zero?

The next two sections will tackle these two questions.

5.2 Initial BFS

Suppose we are trying to solve an LP of the form,

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \\ & x \geq 0. \end{aligned}$$

Note 9. This is not in standard form, but an LP in standard form can easily be converted into one in this form.

Let us see how to find the initial BFS. Remember, we introduced slack variables to make inequalities into equalities.

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax = b \\ & x \geq 0. \end{aligned}$$

If all the right hand sides, b_i, are already positive, then we have the simple BFS with only slack variables. The difficulty lies in the case when $b_i \leq 0$. Suppose the set of equations which have $b_i \leq 0$ is called S. Then introduce an *auxiliary* variable x_s for every $s \in S$.

Look at the following LP,

$$\begin{aligned} \max \quad & -\textstyle\sum_{s \in S} x_s \\ \text{s.t.} \quad & Ax - x_S = b \\ & x \geq 0. \end{aligned}$$

Where x_S is the vector with x_s at s^{th} position and 0 everywhere else. The original LP is feasible iff this LP has optimal value 0.

Exercise 28. Prove it.

The good thing is, this new LP has a very simple BFS to start with. The BFS consists of all the auxiliary variables from set S and all the slack variables from set $[m] - S$.

We can use the method in the previous section to figure out the optimal of this LP. If it is not zero then the original LP does not have a feasible solution. If zero then the BFS of the new LP is the BFS of the original LP (Why??).

Example Suppose the LP given is,

$$\begin{aligned} \max \quad & -2x_1 - 3x_2 \\ \text{s.t.} \quad & x_1 - 3x_2 + 2x_3 \leq 3 \\ & x_1 - 2x_2 \leq -2 \\ & x \geq 0. \end{aligned}$$

We introduce the slack variables,

$$\begin{aligned} \max \quad & -2x_1 - 3x_2 \\ \text{s.t.} \quad & x_1 - 3x_2 + 2x_3 + x_4 = 3 \\ & x_1 - 2x_2 + x_5 = -2 \\ & x \geq 0. \end{aligned} \tag{1}$$

Note 10. Here the slack variables will not give us a BFS like last lecture's example. The problem is one of the b_i's are negative.

After introducing the auxiliary variables,

$$\begin{aligned} \max \quad & -x_6 \\ \text{s.t.} \quad & x_1 - 3x_2 + 2x_3 + x_4 = 3 \\ & x_1 - 2x_2 + x_5 - x_6 = -2 \\ & x \geq 0. \end{aligned}$$

The BFS here is $x_6 = 2, x_4 = 3$ and rest are zero. Changing the equations with respect to the set of non-basic variables,

$$\begin{aligned} \max \quad & -2 - x_1 + 2x_2 - x_5 \\ \text{s.t.} \quad & x_6 = 2 + x_1 - 2x_2 + x_5 \\ & x_4 = 3 - x_1 + 3x_2 - 2x_3 \\ & x \geq 0. \end{aligned}$$

We can increase the variable x_2 up to 1. The new BFS is $x_2 = 1, x_4 = 6$. The new set of equations,

$$\begin{aligned} \max \quad & 0 - x_6 \\ \text{s.t.} \quad & x_2 = 1 + \tfrac{1}{2}x_1 - \tfrac{1}{2}x_6 + \tfrac{1}{2}x_5 \\ & x_4 = 6 + \tfrac{1}{2}x_1 - \tfrac{3}{2}x_6 + \tfrac{3}{2}x_5 - 2x_3 \\ & x \geq 0. \end{aligned}$$

This is the optimal and we have the BFS for the starting LP 1.

6 Degeneracy

Degeneracy arises when there are multiple BFS with the same objective value. The problem is, we can keep circling between these BFS's without ever reaching the optimal vertex. Let us see, how this manifests in our description of simplex algorithm. It is important to understand what is degeneracy, then the solutions are not very difficult.

As an example of degeneracy, suppose we have this set of equations at some stage

$$\begin{aligned} \max \quad & 2 - x_1 + 2x_2 - x_5 \\ \text{s.t.} \quad & x_6 = 2 + x_1 - 2x_2 + x_5 \\ & x_4 = 0 - x_1 - 2x_2 - 2x_3 \\ & x \geq 0. \end{aligned}$$

It is clear that we want to increase x_2, but increasing it even a little bit will make x_4 negative. Hence we can switch x_2 to x_4 in the basic set without changing the solution. What happens if this creates a cycle? This is called degeneracy.

More formally, suppose at any instance the LP is in form,

$$\begin{aligned} \max \quad & C' + \textstyle\sum_{j \in \mathcal{N}} c'_j x_j \\ \text{s.t.} \quad & \forall\ i \in \mathcal{B} x_i = b'_i - \textstyle\sum_{j \in \mathcal{N}} a'_{ij} x_j \\ & x \geq 0. \end{aligned}$$

The degeneracy arises if one of the value of b'_i is zero. Not every such case when b'_i is zero is a problem. The only troubling one is when b'_i corresponding to the leaving variable is zero and the the coefficient of entering variable in that equation is negative.

Exercise 29. Show that in this case, we have at least two BFS's with the same objective value.

Unfortunately there are known examples where the algorithm we have described, where we choose the c'_j with maximum value, will get stuck into a cycle.

Fortunately, the trick to avoid degeneracy is not that difficult and is called *Bland's rule*. Remember that any stage we have a set of choices for the leaving and entering variable (e.g., the set of all j's, s.t., $c'_j \geq 0$, for entering variable).

Definition 3. *Bland's rule: Impose an order on the variables, i.e., call them $x_1, x_2, \cdots, x_n$. For the choice of leaving and entering variable, always choose the one with the least index from the set of choices.*

It can be shown that using Bland's rule, the simplex algorithm will never cycle and hence finish in finite time. We will only give an idea of the proof. The full proof can be found at `https://www.math.ubc.ca/~anstee/math340/340blandsrule.pdf` .

For the sake of contradiction, let there be a cycle. Every basic solution in this cycle will have the same objective value (condition for degeneracy).

Say x_t is the variable with the largest index from the list of leaving variables (or entering variables) in the cycle. Suppose B is the BFS where x_t leaves and x_s enters, we know $s < t$. Also, let B' be the BFS where x_t enters the basic solution again.

Let the equations when B is basic to be,

$$\begin{array}{c} \max \quad C + \sum_{j \notin B} c_j x_j \\ \text{s.t.} \quad x_i = b_i - \sum_{j \notin B} a_{ij} x_j \quad \forall i \in B \\ x \geq 0. \end{array}$$

Let the equations when B' is basic to be,

$$\begin{array}{c} \max \quad C' + \sum_{j \notin B'} c'_j x_j \\ \text{s.t.} \quad x_i = b'_i - \sum_{j \notin B'} a_{ij} x_j \quad \forall i \in B' \\ x \geq 0. \end{array}$$

We will not worry about the constraints for B'. Let $x_s = \delta$ and $x_i = b_i - a_{is}\delta$ for all $i \in B$ (0 otherwise) be a solution for the equations when B is basic. We do NOT claim that $x \geq 0$, just that x satisfies the set of equations.

Since the objective function with respect to B' is just the objective function of B added with multiples of constraints with respect to B, the objective function for B and B' should be same for this paticular solution.

Comparing the objective function, we get an $r \in B$ such that $c'_r a_{rs} < 0$ ($r \notin B'$ otherwise $c'_r = 0$). Since $c'_t \geq 0$ and $a_{ts} \geq 0$ (why), $r \neq t$. The variable t entered at B' and not r, implies $c'_r \leq 0$, so $a_{rs} > 0$. In that case, r should have left B and not t, giving contradiction.

7 Assignment

Exercise 30. Show that the vector space associated with an affine set does not depend upon the point chosen.

Exercise 31. Prove that the line (in $\mathbb{R}^2$) between x_1 and x_2 is the set of all points p, s.t.,

$$a^T p = a^T x_1 = a^T x_2.$$

Here a is the unit vector perpendicular to $x_1 - x_2$.

Exercise 32. We have given combination interpretation (affine, convex, conic) of polytope, cones and bounded polytopes. Can you give similar interpretation for hyperplanes and halfspaces?

Exercise 33. Convex hulls and cones are closely related. Take x_i's as row vectors. Prove,

$$x \in Conv(x_1, x_2, \cdots, x_k) \Leftrightarrow (x, 1) \in Cone((x_1, 1), \cdots, (x_k, 1))$$

Exercise 34. Given a convex set S, show that the set

$$T = \{x : Ax + b = y,\ y \in S\}$$

is also convex.

Exercise 35. A *Minkowski sum* of two sets S_1, S_2 is the set formed by taking all possible sums such that first vector is from S_1 and second vector is from S_2.

$$S_1 + S_2 = \{x : x = x_1 + x_2, x_1 \in S_1, x_2 \in S_2\}.$$

Prove that the Minkowski sum of two convex sets is convex. You can prove a stronger result too,

$$Conv(S_1 + S_2) = Conv(S_1) + Conv(S_2).$$

Where $Conv(S)$ takes the convex hull of set S.