Advanced Optimization: Theory and Applications

Rishabh Singhal, Nishant Sachdeva

February 7, 2021

Contents

hyperref

1 Convex Optimization: Theory - Lecture 1

1.1 Topics to be covered

1. Convex Optimization

2. Non-Convex Optimization

3. Non-Smooth Optimization (non-differentiable)

4. Combinations of (2) and (3)

5. Min-Max Optimization (GAN: generative adversarial networks)

6. Optimization on Manifolds (as domain – some structure)

7. Distributed Optimization (Machine Learning problems with multiple workers, a distributed environment.

1.2 Implicit Function Theorem

Definition 1. ***Implicit*** *relations are of the form* $F(x, y) = 0$ *while* ***Explicit*** *functions are of the form* $y = f(x)$.

Theorem 1.
Given a point (x_0, y_0)*, Implicit function Theorem deals if an implicit relation* $F(x, y)$ *is a function in the local neighbourhood of* (x_0, y_0) *such that* $F(x_0, y_0) = 0$ *and deals with those conditions.*
NOTE: *TBA - These conditions.*

1.3 Gradient and Level Curves

Gradients basically a tool for optimization if the structure is continuous and smooth (that is differentiable).

1. $\nabla z = \nabla f(x, y) = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, ...)$, a vector for 2D function.

2. Points towards the direction of instantaneous increase of the function.

3. It is denoted by $\nabla f(x)$.

Definition 2. ***Level Curve.*** *This is the curve obtained by setting certain heights to the function* $f(x, y)$)*, i.e. curve obtained by plotting* $f(x, y) = t$ *for some fixed* t *belonging to some set. Formally, the level curve of a scalar valued function*

$$f : \mathbb{R}^n \mapsto \mathbb{R}$$

is defined as a set as follows:

$$f_t = \{(x, y)|f(x, y) = t, t \in \mathbb{R}\}$$

Example. *Plot* ***level curves*** *of*

1. $f(x, y) = x^2 + y^2$

2. $f(x, y = x^2 - y^2$

for $t = 0, 1, 2$.

Some notes:

- For level curves there is a need of color coding to differentiate between minima and maxima as they both look similar.

- At a given point, gradient points in the direction perpendicular to level curve. [For proof refer **Implicit Function Theorem** which is skipped for this course].

- The **projection** of Gradient (in case of 3D) on 2D plane vanishes at maxima or minima or we can say that 2D plane (when level curve is considered).

1.4 Gradient Descent

Now, for the gradient descent the algorithm goes like, moving from one point to another so as to follow a direction parallel to opposite of what gradient directs to. Basically,

$$x^{i+1} \rightarrow x^i - t \times \nabla f(x^i) \tag{1}$$

for $i = 0, 1, 2, ...$ where t is the learning rate/step length (which shows how far to move). And x^0 is initialized with some random point.

Next question is, when to stop? Either when $||\nabla f(x^i)|| \approx 0$ or $||x^i - x^{i-1}|| \approx 0$

Similarly for gradient ascent. (to reach local maxima) equation can be modified as follows

$$x^{i+1} \rightarrow x^i + t \times \nabla f(x^i) \tag{2}$$

Now, these algorithms works well on a bowl shape curve. Which are nothing but convex curves and hence the name convex optimization.

1.5 Convex Set and Convex Functions

Definition 3. ***Convex Set.*** *A set is convex if* ***all convex combinations*** *lie in the set*

$$\sum \theta_i x_i = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots, \ \sum \theta_i = 1, \ \theta_i \geq 0, \forall i$$

Another definition for convex set is as follows

Definition 4. ***Convex Set.*** *A set is convex if every line between two points stays in the set*

$$\theta x_1 + (1 - \theta) x_2, \ 0 \leq \theta \leq 1$$

Here, x_2 can be seen as the base point and then with θ length going towards $(x_1 - x_2)$ so θ must be in between 0 and 1, otherwise if it is greater than 1 or less than 0 then the final vector will will lie outside.

Why this works? The implication from previous definition to this definition is implicit. The previous definition can be constructed from this definition by allowing more variables (with constraints) and defining for example line between the third point, and the arbitrary point between two given points. And using mathematical induction, the general definition can be obtained. So, these two are equivalent.

Definition 5. ***Convexification*** *of a set can be done by arbitrary choosing two points and adding all the points lying in between these two points (if they are not already present). This need to be done until the set becomes convex.*

Some examples of convex set,

1. **Hyperplane:** $\{x | a^T x = b\}$

$$\begin{aligned} b = a^T x = \vec{a} \cdot \vec{x} &= (||\vec{a}||) \cdot (||\vec{x}|| \cos \theta) \\ &= \text{const} \cdot \text{shadow of x on a} \\ \text{shadow of x on a} &= \frac{\text{const}}{b} = \text{const2} = k \end{aligned}$$

   Hence, shadow of $\vec{x}$ on $\vec{a}$ should have constant length defined by k, which give rise to the hyperplane.

2. **Half Space:** $\{x | a^T x \geq b\}$ or $\{x | a^T (x - x_0) \geq 0\}, \ b = a^T x_0$

3. **Norm Balls:** Consider the norm balls (can be compared to filled balloon in case of $l = 2$).

$$\{x \in \mathbb{R}^n |\ ||x||_l = 1\}$$

these are not convex.

- Norms are equivalent for all l **it can be infinite also**.

4. **Sphere (Balls):** $\{x \in \mathbb{R}^n |\ ||x - x_0||_l \leq 1\}$ For $l = 1, 2, 3, \cdots$ sphere is convex.

5. **Ellipsoid:** $\{x|\ (x - x_c)^T P^{-1}(x - x_c) \leq 1\}$

- Here P is symmetric positive definite **(that means all it's eigenvalues are positive)**.
- Positive definite matrix is a matrix M, such that $z^* M z > 0\ \forall z$ where z^* is complex conjugate transpose and M is **hermitian** (i.e. $M^* = M$).
- Matrix P determines how far the ellipsoid extends in every direction from center x_c.
- The length of semi-axes and major-axis of ellipsoid are given by minimum and maximum of $\sqrt{\lambda_i}$ respectively, where λ_i are eigenvalues of $P\ \forall i$.
- A ball (sphere, **note** it's not norm ball) is an ellipsoid with $P = r^2 I$.
- Another representation of Ellipsoid

$$\{x_c + Au|\ ||u||_2 \leq 1\},\ \text{A is SPD}$$

6. The **norm cone** associated with a norm $||.||_l$ is the set

$$C = \{(x, t)|\ ||x||_l \leq t\} \subseteq \mathbb{R}^{n+1}$$

where $x \in \mathbb{R}^n$

- alternative is the set of points $(x, y, z, ..., t)$ $x^2 + y^2 + z^2 + ... \leq t$
- Other names of second order cone (l = 2): **quadratic cone, Lorentz-cone, ice-cream cone**
- **Note:** it is not convex

7. **Polyhedron** is defined as a set of finite number of equalities and inequalities:

$$P = \{x|\ {a_j}^T x \leq b_j,\ j = 1, \cdots, m,\ c_j^T x = d_j,\ j = 1, 2, \cdots, p\}$$

- Intersection of finite number of halfspaces and hyperplanes.
- A polyhedron can be more compactly represented as

$$P = \{x|Ax \leq b, Cx = d\}$$

  where A and C are a matrix of vectors
- Example: **non-negative orthant**

$$\mathbb{R}^n = \{x \in \mathbb{R}^n |\ \vec{x} \geq \vec{0}\}$$

  Note here the rule used for sequencing is comparing each element of x with 0.
- A bounded polyhedron is called **polytope**.

8. **Siplex** is the convex set of $k + 1$ affinely independent vectors.

Definition 6. *A set of $k+1$ vectors is said to be the set of affinely independent vectors if $v_1 - v_0, v_2 - v_0, ..., v_k - v_0$ are linearly independent and affine dimension is k.*

- Unit simplex: vectors = $\{0, e_1, ..., e_k | \ e_i \in \mathbb{R}^n\}$
- Probability simplex: vectors = $\{e_1, ..., e_k | \ e_i \in \mathbb{R}^n\}$

9. **Cone** A set C is called cone if for every $x \in C$, and $\theta \geq 0$, we have $\theta x \in C$.
10. **Convex Cone** A set C is called convex cone if it is convex and a cone.
11. S^n is the set of symmmetric matrix.
12. S_+^n is the set of semi-positive definite symmetric matrix.
    - $det(A) = \Pi_i \lambda_i$ where λ_i are eigen values of A.
13. S_{++}^n is the set of positive definite symmetric matrix.

Note: To graph the set of matrix, just take any general representation of symmetric matrix and impose other conditions on this matrix. The set of inequalities will determine the graph together.

Definition 7. ***Conic Combination*** *A point of the form $\theta_1 x_1 + \theta_2 x_2 + + \theta_k x_k$ with $\theta_1, \theta_2, ..., \theta_k \geq 0$ is called the conic combination of $x_1, x_2, ..., x_k$.*

Definition 8. ***Conic Hull*** *of a set of vectors is the set of all conic combinations of those vectors.*

Theorem 2. *Intersection between convex sets is* ***closed.***
Note: *This is trivial.*

2 Convex Optimization: Theory - Lecture 2

2.1 Convex Sets

Theorem 3.

The semi positive definite cone is convex.

Proof. Semi Positive definite cone is an intersection of the following set.

$$S_+^n = \bigcap_{z \neq 0} \{X \in S^n | \ z^T X z \geq 0\}$$

So , if we can prove that this set is convex, then using fact 1, we will have proved that the positive semi definite cone is also convex. ...(1)

To prove that the $z^T X z$ is convex, we first establish that the set is , in fact, linear. A function $f(x)$ is linear if

1. $f(x+y) = f(x) + f(y)$
2. $f(cx) = cf(x)$

Here , we can see that $z^T X z$ follows both the conditions of being linear. (trivial proof) ...(2)

So, from (1) and (2) , we can say that the set $z^t X z$ is convex. Further, using Fact 1 (the intersection of convex sets is convex) , we conclude that Positive Semi-Definite Matrices are a convex set.

□

2.2 Affine Functions

Definition 9. *A function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is an* ***affine*** *function if it is a sum of a linear function and a constant i.e. of the form*

$$f(x) = Ax + B,\ A \in \mathbb{R}^{m \times n},\ \mathbb{R}^m$$

Theorem 4. *If $f(x)$ is an affine function and S is a convex set then $f(S)$ and $f^{-1}(S)$ are convex sets too.*

For example,

Theorem 5. *Polyhedron is a convex set.*

$$P = \{x|\ Ax \leq b, Cx = d\}$$

Proof. Consider the set

$$S = \{(x, 0)|\ x \geq 0\},\ \text{non-negative orthant}$$

Consider the affine function

$$f(x) = (b - Ax, d - Cx)$$

It can be seen that $f^{-1}(S) = P$ □

Definition 10. ***Supremum.*** *Least of all upper bounds, it can be that it does not exist in the set. Also, when the set might not have min/max, it can have infimum/supremum.*

Notes:

- Without upper bound $\mapsto$ unbounded above.

Definition 11. ***Supporting Hyperplane theorem.*** *Suppose C and D be convex sets, such that $C \cap D = \emptyset$, then there always exists atleast one hyperplane (with $a > 0$) separating the two sets i.e. sets C and D lies on opposite side of the hyperplane.*

2.3 Interior Points, Closed Sets, Open Sets

Definition 12. ***Interior point.*** *An element $x \in C \subseteq R^n$ is called interior if there exists an $\epsilon > 0$ such that,*

$$\{y|\ ||y - x||_2 \leq \epsilon\} \subseteq C$$

Definition 13. ***Open Sets.*** *A set with all points as interior points.*

Definition 14. *Set of all interior points in a set C is marked as $intC$.*

Definition 15. ***Closed Sets.*** *A set C is closed if the complement $\mathbb{R}^n \setminus C$ is open.*

Definition 16. ***Closure.*** *A closure of set C is the set with it's boundary points.*

Definition 17. ***Boundary.*** *A point whose ball's intersection is non-empty with both C and C^c, it can also be not a part of the set itself.*

Examples:

- $\{\sin x \mid 0 \leq x \leq 2\pi\} = [-1, 1]$, here -1 and 1 are boundaries and the set is closed set.
- $\{(\frac{1}{m}, \frac{1}{n}) \mid m, n \in \mathbb{R}\}$

3 Convex Optimization: Theory - Lecture 3

3.1 Supporting Hyperplane

Definition 18. ***Supporting hyperplane.*** *Let $C \subseteq \mathbb{R}^n$, and x_0 is a boundary point. If $a \neq 0$ satisfies $a^T x \leq a^T x_0$ for all $x \in C$ then the hyperplane $\{a^T x = a^T x_0\}$ is supporting hyperplane. It always go out of the surface.*

Definition 19. ***Dual Cone.*** *Let K be a cone. Then*

$$K^* = \{y|\ x^T x \geq 0\ \forall x \in K\}$$

is called the dual cone of K. Note: It is convex regardless of the convexity of K and it's a convex cone.

It can be useful if the given cone is not convex and we can move to dual cone to optimize and come back to the previous cone domain.

- Dual cone contains set of all y such that $-y$ is the normal to supporting hyperplane. Think about proof? Trivial.

3.2 Convex Functions

Definition 20. *A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is convex if* ***dom*** *f is a convex set and if for all $x, y \in f$, and θ with $0 \leq \theta \leq 1$, we have*

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

Definition 21. ***Strict function.*** *if strict inequality holds for $x \neq y$ and $0 < \theta < 1$.*

- A function is convex iff it is convex when **restricted** to any line that intersects it's domain.
- f is convex iff consider a function $g(t) = f(x + tv)$ where $x + tv \in dom\ f$ is convex.

3.3 Extended Value Function

Definition 22. ***Extended value function.*** *If f is convex then the extended value function $\tilde{f} : \mathbb{R}^n \mapsto \mathbb{R} \cup \infty$ by*

$$\tilde{f}(x) = x \in dom\ f ? f(x) : \infty$$

Note: Convexity is not violated.

Definition 23. ***Indicator Function.*** *Let $C \subseteq \mathbb{R}^n$ be a convex set, then the indicator function is the function which is 0 in the domain C while it is ∞ for $x \notin C\ \forall x \in \mathbb{R}^n$.*

Note:

- To optimise f over some convex set C, it is same as optimising over the function $f + \tilde{I}_C$.
- Set C is convex if I_C is convex function.

3.4 Convexity - First and Second Order Condition

Definition 24. ***First Order Condition.*** *If f is differentiable, then f is convex if and only if domain of f is convex and*

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

holds for all $x, y \in dom\ f$.
NOTE:

- *connection with taylor expansion can be seen.*
- ***Global under-estimator.***
- *Also acts like a supporting hyperplane.*
- *Suppose $f(x) = 0$ then $f(y) \geq f(x)\ \forall y$.‘*

Definition 25. ***Second Order Condition.*** *If f is twice differentiable, then f is convex if and only if domain of f is convex and*

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

NOTE:

- *Hessian ∇^2*
- *$\nabla^2 f(x)$ is Positive semi definite matrix*
- *Curve with positive curvature*

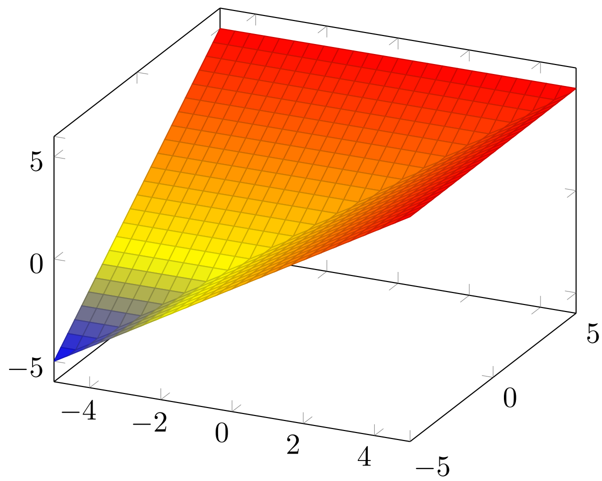
For both of the above conditions, the *dom f* should be convex. For example consider a counter example,

$$f(x) = \nVdash x^2,\ dom\ f = \{x \in \mathbb{R} |\ x \neq 0\}$$

this satisfies $f''(x) > 0\ \forall x \in dom\ f$ but it is not convex function.

$\mathbb{R}_{++}$ is set of strictly positive real numbers.

- **Norm** is a convex function. Hint: Triangle Inequality



- $f(x) = \max x_1, x_2, ..., x_n$ for $x \in \mathbb{R}^n$ is convex.

NOTE: For positive definite square root exist. Also $det(A \times B) = det(A) \times det(B)$

Example. ***Log-determinant*** *$f(X) = .\log det\ X, X > 0$ is a concave function.*

Proof.

- Consider arbitrary line $X = Z + tV,\ Z, V \in \mathbb{S}^n$ and positive definite.
- Consider $g(t) = f(Z + tV), t \in [0, 1]$
- We have

$$\begin{aligned} g(t) &= \log det(Z + tV) \\ &= \log det(Z^{1/2}(I + tZ^{-1/2}VZ^{1/2})Z^{1/2} \\ &= \sum_{i=1}^{n} \log 1 + t\lambda_i + \log det(Z) \end{aligned}$$

  Where $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of $Z^{-1/2}VZ^{-1/2}$.

□

3.5 Types of Graphs

Definition 26. ***Epigraph*** *of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is defined as*

$$epi\ f = \{(x, t)|\ x \in dom\ f, f(x) \leq t\}$$

A function is **convex** iff it's epigraph is convex.

Definition 27. ***Hypograph*** *of a function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is defined as*

$$epi\ f = \{(x, t)|\ x \in dom\ f, f(x) \geq t\}$$

A function is **concave** iff it's hypograph is concave.

4 Convex Optimization: Theory - Lecture 4

Definition 28. ***Jensen's Inequality.*** *For a convex function (which is continuous),*

$$f(\theta_1 x_1 + ... + \theta_k x_k) \leq \theta_1 f(x_1) + ... + \theta_k f(x_k),$$

given $\theta_i \geq 0\ \forall i$ and $\sum \theta_i = 1$.

Definition 29. ***Holder's inequality.*** *Let $p > 1, 1/p + 1/q = 1,\ x, y \in \mathbb{R}^n$*

$$\sum_{i=1}^{n} x_i y_i \leq (\sum_i = 1^n |x_i|^p)^{1/q} (\sum_i = 1^n |x_i|^q)^{1/q}$$

- Weighted combination of convex function is convex.
- Composition of affine functions also give convex under some conditions.

4.1 Conjugate Function

- Dual and **Primal** functions.

Definition 30. ***Conjugte Function.*** *Let $f : R^n \mathbb{R}$. Then the function $f^* : \mathbb{R}^n \mapsto \mathbb{R}$ defined as*

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

is called the conjugate of the function f. The domain of the conjugate function consists of $y \in \mathbb{R}^n$ for which the sup is finite, i.e., the difference $y^T x - f(x)$ is bounded above. For $1D$, right side function is called Fenchel/Legendre's transform.

4.2 Convex Optimization

Definition 31. ***Optimization Problem. (Standard Form)***

$$\begin{aligned} \textit{minimize } & f_0(x) \\ \textit{subject to } & f_i(x) \leq 0 \\ & h_i(x) = 0 \end{aligned}$$

Note, the number of h_i and f_i can be different or zero.

5 Convex Optimization: Theory - Lecture 5

- Two optimization problems are equivalent if the solution of one can be obtained from other.
- **Slack variable** can also be introduced with non-negativity constraint to transform inequality to equality constraint.
- Non-negative constraints are easier to deal with?
-

Definition 32. ***Convex Optimization Problem. (Standard Form)***

$$\begin{aligned} \textit{minimize } & f_0(x) \\ \textit{subject to } & f_i(x) \leq 0 \\ & h_i(x) = 0 \end{aligned}$$

$f_0(x)$ is convex, f_i should be convex and $h_i(x)$ should be ***affine.***

- Note to check double derivative of a vector $x \in \mathbb{R}^2$, there is a need to use hessian $\mathbb{H}$.
- See if optimization problem can be converted into convex optimization problem (which can be solved using some current library).

5.1 Optimization and Duality

- Local minima ==¿ Global minima (for convex function)
- If the function is differentiable and x is optimal point then

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

  This can be obtained from the Word of first order definition (or property?) of convex functions. Also it can be reduced to **(in case of constrained convex optimization problem)**

  $$\nabla f_0(x) = 0$$

- A linear function (which is always greater than 0) can be only 0.
- $\mathbb{R}(A)$ is the set of vectors Ax.

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

because $\nabla f_0^T(x)v$ is linear in v, therefore $\nabla f_0^T(x)v = 0$ where $v \in \mathbb{N}(A)$

Because, consider a linear function $g(x)$ in x, then as $g(x)$ is linear therefore

$$g(\alpha x) = \alpha g(x)$$

now, if $g(x) \geq 0$ for all x, then $g(\alpha x) \geq 0 \implies \alpha g(x) \geq 0$, here α is a constant which can be both negative and positive; and the only number to which if you multiply a negative or a positive number and the output must be greater than 0 is **0**.

$$\begin{aligned} \nabla f_0^T(x)v &= 0 \\ \nabla f_0 &\perp v \end{aligned}$$

Therefore,

$$\begin{aligned} Av &= 0 \\ (A^T)^T v &= 0 \\ \nabla f_0(x) &\in \mathbb{R}(A^T) \end{aligned}$$

so,

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

which is langrange condition

- F can be seen as the matrix which encompasses the null space of A. (which can be used to eliminate equality condition) where columns of F is the basis of null space of A.

5.2 Different types of programs

- Linear Program
- Quadratic Program (with symmetric and positive definite P
- QCQP (quadratic constraint – quadratic program)
- Second order cone programming
- Robust linear program