MIT 6.7220 Nonlinear Optimization(2024)

Ian Poon

October 2024

Selected theorems from professor `Gabriele Farina's` lecture notes for `MIT 6.7220`. This is my first graduate theoretical computer science module :). Not finished yet tho...will return when I actually need to apply covergence properties in ML

Contents

1 introduction(1)

1.1 optimization problems and some prelminary considerations

In general an optimization problem has the form

$$\min_x f(x)$$

for $x \in \Omega$ where

- $f(x)$ is called the **objective function**
- the entries of the vector x are called **optimization variables**
- The set Ω is called the **feasible set** or **constraint** set

In particular for this set of notes, we are interested in optimization problems that satisfy the following assumptions

- assume that Ω is a *finite* embedding in a euclidean space
- the objective function $f(x)$ is a *continuous* real-valued function
- our problem has *functional constraints* defined as follows

$$\Omega := \left\{ x \in E \;\middle|\; \begin{array}{ll} g_i(x) \le 0, & i = 1, \dots, r, \\ h_j(x) = 0, & j = 1, \dots, m. \end{array} \right\}.$$

Theorem 1 (Weierstrass theorem)
Let $f : \Omega \to \mathbb{R}$ be a *continuous* function defined on a nonempty and compact set Ω. Then there exists a minimizer $x^* \in \Omega$ of f on Ω that is

$$f(x^*) \le f(x)$$

Proof. This is a basic analysis result. Refer back to your notes on `Rudin` if necessary. The above is essentially the infinum of the function.

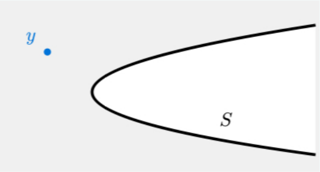
While Weierstrass theorem is a pretty universal tool, many optimization problems however do not start off with a bounded feasible set Ω. For example,

Example 2

Let $y \in \mathbb{R}^n$ be a point and $S \subseteq \mathbb{R}^n$ be a *closed* but not necessarily bounded set. Prove that there always exists a projection point

$$\min_x ||x - y||$$

where $x \in S$.

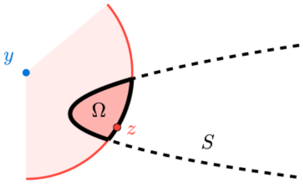


In other words, a point in S such that the "distance" between x, y is a minimum.

Solution. We know that for sure if a minimizer exists it must live in the intersection

$$\Omega = S \cap \overline{B}_{||z-y||}(y)$$

where $\overline{B}_{||z-y||}(y)$ represents the closed ball centered at y with radius $||z - y||$. Well if you define distance you can obviously define a closed ball neighbourhood.



Now because a finite intersection of closed sets is closed and Ω is non empty as it contains z it follows that Ω is compact. Therefore by **wierstrass theorem** above, the proposition follows □

More generally we have

Theorem 3 (Weirstrass theorem for compact sublevel sets)

Let f be a continuous function defined on a set S. If f has a nonempty and compact **sublevel** set, that is, there exist $a \in \mathbb{R}$ such that

$$\{x \in S, f(x) \leq a\}$$

is nonempty and bounded, then f has a minimizing point in S

Proof. This is clearly just a condition that directly allows for the application of Weierstrass theorem □

Remark 4. *In 2, our function was $||x - y||$ and recall from `functional analysis` that norms are continuous and the sublevel set was Ω. This is because $f(x) \leq a$ represents the ball in our example and it has to intersect S.*

Let us consider how we can formulate a nonlinear optimization problem. We will explore "nonlinear" more in future sections.

Example 5 (MAX-CUT as a nonlinear optimization problem)

MAX-CUT is one of the most fundamental problems in theoretical computer science and discrete optimization. It is defined by

- Let $G = (V, E)$ be a graph where $V = \{1, 2, \dots, n\}$ is the set of vertices
- we want to partition V into two sets S and V/S such that the number of edges going between S and V/S is maximized

Solution. First we try to pose this as an optimization problem. Let's assign to each node $v \in V$ a variable $x_v \in \{-1, +1\}$ where

- if $v \in S$ then $x_v = +1$
- if $v \in V/S$ then $x_v = -1$

Consider an arbitrary edge $(u, v) \in E$. Then $x_u \cdot x_v = -1$ if and only if we are connecting one node in S to one node in V/S. We call such edges the **cut edge**.Otherwise if connecting within the same set, we will have $x_u \cdot x_v = +1$. Hence we can count all edges in the cut by considering the objective function

$$h(x) = \sum_{(u,v)\in E} \frac{1 - x_u \cdot x_v}{2}$$

see that obviously $\frac{1-x_u \cdot x_v}{2} = 1$ if and only if it corresponds too a cut edge otherwise the summand is zero. Hence we were able to pose this situation as *discrete* optimization problem where we have to find

$$\max_x h(x)$$

where $x_v \in \{-1, +1\}, \forall v \in V$

2 first order optimality conditions(2)

2.1 unconstrained optimization

Basic elementary multi-variable calculus, essentially $\nabla f(x) = 0$ which will correspond to the values at the critical points.

But for the sake of completeness to demonstrate how this result follows from the familiar single variable calculus result(that says the critical points of first derivative equal zero) Let x be a minimizer of the function $f : \mathbb{R}^n \to \mathbb{R}$ along the direction d. Clearly we have $f(x + t \cdot d) \geq f(x)$ for all $t \geq 0$ since x is a minimizer. Therefore the directional derivative $f'(x; d)$ of f at x along direction d which can be expressed as

$$f'(x; d) = \lim_{t\downarrow 0} \frac{f(x + t \cdot d) - f(x)}{t} \geq 0$$

where the equality follows because the limit of nonnative sequence must be nonnegative. By defintion of gradient we know that $f'(x; d) = \langle \nabla f(x), d\rangle$ so the previous inquality can be rewritten as

$$\langle \nabla f(x), d\rangle \geq 0, \quad \forall d \in \mathbb{R}^n$$

Because the above inquality must hold for all directions $d \in \mathbb{R}^n$ this includes $d = -\nabla f(x)$ too therefore

$$-\,||\nabla f(x)||^2 \geq 0$$

but clearly $-\,||\nabla f(x)||^2 < 0$ so only the equality case is valid so we have $||\nabla f(x)||^2 = 0$. Then by the definiteness of norms this implies

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

as desired.

2.2 constrained optimization

So you might wonder what is the difference between constrained and unconstrained optimization?

Fact 6
The main difference with the constrained case is that in a constrained set we might be limited in the choices of available directions d along which we can approach x while remaining in the set.

Explicitly we write

$$\langle \nabla f(x), d \rangle \geq 0$$

for all $d \in \mathbb{R}^n$ that remain in Ω from x

Definition 7 (Star convexity at x)
A set $\Omega \subseteq \mathbb{R}^n$ is said to be **star convex** at a point $x \in \Omega$ if for all $y \in \Omega$ the entire segment from x to y is contrained in Ω. In symbols this means

$$x + t \cdot (y - x) \in \Omega \quad \forall t \in [0,1]$$

Definition 8 (Convex set)
A set Ω is convex if it is star convex at *all* of its points $x \in \Omega$. In symbols this means

$$x + t \cdot (y - x) \in \Omega \quad \forall x, y \in \Omega \text{ and } \forall t \in [0,1]$$

In other words all segments formed between any two point $x, y \in \Omega$ are entirely contained in Ω

Theorem 9 (First order necessity optimality condition for a convex feasible set)
Let $\Omega \subseteq \mathbb{R}^n$ be convex and $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. For a point $x \in \Omega$ to be a minimizer of f over Ω it is necessary that

$$\langle \nabla f(x), y - x \rangle \geq 0 \quad \forall y \in \Omega$$

Proof. No brainier. $y - x$ defines a direction d. Then just apply the arguments above

The condition established in 9 has some geometric interpretation.

Definition 10 (Normal Cone)
Let $\Omega \subseteq \mathbb{R}^n$ be convex and let $x \in \Omega$. The **normal cone** to Ω at x denoted $\mathcal{N}_\Omega(x)$ is defined as the set

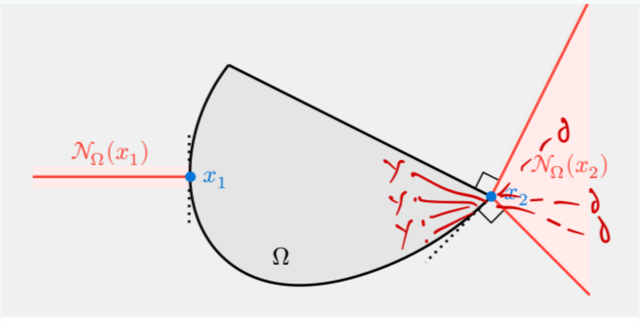
$$\mathcal{N}_\Omega(x) = \{d \in \mathbb{R}^n : \langle d, y - x\rangle \leq 0 \ \forall y \in \Omega\}$$

With this definition the first order necessary optimally condition for x given in 9 can be equivalently written as

$$\boxed{-\nabla f(x) \in \mathcal{N}_\Omega(x)}$$

by simply substituting $d = -\nabla f(x)$. This is a nice result, essentially for x to be a minimizer its gradient must lie in the normal cone to Ω at x.

Example 11
Consider a few examples of normal cones on a convex set here

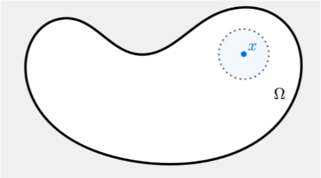


This makes sense if you recall elementary trigonometry that $\langle d, y - x\rangle \leq 0$ implies the angle between d and $y - x$ is obtuse/bigger than 90°. Alternatively recall the definition of dot product and just think of the graph of the cosine function.

2.3 normal cones to some notable sets and applications

Consider the 2 examples below. Honestly you can probably visualize the proposition without the proof intuitively. But for rigour we will quickly run through the proofs.

Example 12 (Normal Cone at an interior point)
What is the normal cone $\mathcal{N}_\Omega(x)$ of a point x in the interior of the feasible set Ω?



Proof. Consider any direction $d \neq 0$. In the interior neighbourhood centered at x it is clear any there exists a point that can be represented by $x + \delta d$ for some sufficiently small $\delta > 0$. However that implies there exists a point such that

$$\langle d, x + \delta d - x\rangle = \delta \, ||d||^2 > 0$$

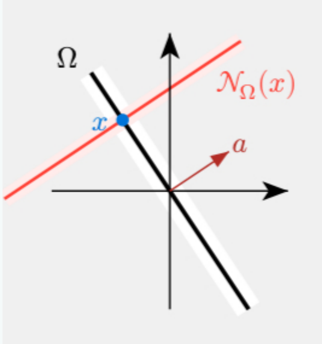
which violates the necessary condition that $\nabla f(x) = 0$ when we let $d = -\nabla f(x)$ like previously. Therefore the normal cone only contains the zero vector

$$\mathcal{N}_\Omega(x) = \{0\}$$

Example 13 (Normal Cone to a hyperplane)
Consider a hyperplane

$$\Omega = \{y \in \mathbb{R}^n : \langle a, y \rangle = 0\}$$

where $a \in \mathbb{R}^n$, $a \neq 0$ and a point $x \in \Omega$.



Show that

$$\mathcal{N}_\Omega(x) = \text{span}\,\{a\} = \{\lambda \cdot a : \lambda \in \mathbb{R}\}$$

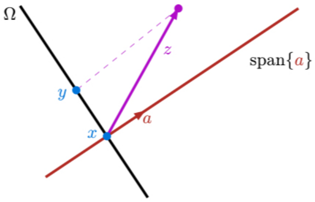
Remark 14. *Intuitively you can see why this must be true. Suppose the normal cone is not aligned with the normal,at any $x \in \Omega$ then on one side(to its left or right) it will be less than 90° away from a direction $y - x$(which violates the condition to be a normal cone). But to show this rigorously, consider the following*

Solution. To show that $\text{span}\,\{a\} \subseteq \mathcal{N}_\Omega(x)$ simply consider

$$\langle z, y - x \rangle = \langle \lambda \cdot a, y - x \rangle = \lambda \cdot \langle a, y \rangle - \lambda \cdot \langle a, x \rangle = 0 - 0 \leq 0$$

Which follows by definition when $y \in \Omega$ and that clearly from the diagram $x \in \Omega$ too.

From the other direction to show that $\mathcal{N}_\Omega(x) \subseteq \text{span}\,\{a\}$. It's equivalent contra-positive statement is that $\text{span}\,\{a\}^c \subseteq N_\Omega^c$. Suppose for the sake of contradiction there exists $z \in \text{span}\,\{a\}^c$ that is a member of $\mathcal{N}_\Omega$. Now consider



That implies we may write z as

$$z = (y - x) + k \cdot a$$

Then subbing this into

$$\langle z, y - x \rangle = \langle (y - x) + k \cdot a, y - x \rangle = ||y - x||^2 + k \cdot \langle a, y - x \rangle = ||y - x||^2$$

because $y \neq 0$ this means $||y - x||^2 > 0$ and so it cannot be in $\mathcal{N}_\Omega$ so we have a contradiction.

Corollary 15

Consider

$$\Omega = \{y \in \mathbb{R}^n : Ay = b\}$$

where $A = a^T$. Then

$$\mathcal{N}_\Omega(x) = \text{span}(A^T)$$

Proof. Consider that the above corresponds to the case when $b = 0$ and that the results hold regardless of the value of b since

$$\langle z, y - x \rangle = \langle \lambda \cdot a, y - x \rangle = \lambda \cdot \langle a, y \rangle - \lambda \cdot \langle a, x \rangle = b - b \leq 0$$

Next notice that $a^T y$ is simply the euclidean inner product/dot product $\langle a, y \rangle$. □

This result immediately implies for x to be a minimizer of $f(x)$ given that $Ax = b$ we must have $-\nabla f(x) = A^T \lambda$ for some $\lambda \in \mathbb{R}^d$. The entries of λ are examples of what we call **Lagrange multipliers**

Let us consider an example application of this.

Example 16

Consider the nonempty set $\Omega = \{x \in \mathbb{R}^n : Ax = b\}$ where $A \in \mathbb{R}^{d \times n}$ is such that AA^T is invertible. Prove that the euclidean projection x of a point z onto Ω that is the solution to

$$\min_x \frac{1}{2} ||x - z||_2^2$$

where $x \in \Omega$ is given by

$$x = z - A^T (AA^T)^{-1} (Az - b)$$

Our objective function is $f(\mathbf{x}) = \frac{1}{2} ||\mathbf{x} - \mathbf{z}||_2^2$. Now since $||\mathbf{x} - \mathbf{z}||_2^2 = \left(\sqrt{(\mathbf{x} - \mathbf{z}) \cdot (\mathbf{x} - \mathbf{z})}\right)^2 = (\mathbf{x} - \mathbf{z}) \cdot (\mathbf{x} - \mathbf{z}) = \sum_i (x_i - z_i)^2$. Then

$$\nabla \frac{1}{2} ||\mathbf{x} - \mathbf{z}||_2^2 = \frac{1}{2} \nabla \sum_i (x_i - z_i)^2 = \frac{1}{2} \sum_i 2(x_i - z_i)\mathbf{e}_i = \mathbf{x} - \mathbf{z}$$

Therefore for x to be a minimizer we require that

$$-\nabla f(x) = -(x - z) \in \mathcal{N}_\Omega(x)$$

But because Ω is such that x satisfies $Ax = b$, from 15 it follows that

$$-(x - z) = A^T \lambda \quad \Rightarrow \quad x = z - A^T \lambda$$

Since $x \in \Omega$ we have $Ax = b$ so subbing our expression from above in we have

$$A(z - A^T \lambda) = b \quad \Rightarrow \quad (AA^T)\lambda = Az - b$$

Solving for λ and plugging it back into $x - A^T \lambda$ yields the result □

2.4 Application: derivation of linear programming duality

Consider the linear program

$$\max_x f(x) = c^T x$$

where $Ax \leq b$ and $x \in \mathbb{R}^n$. We call this the **primal problem**(P). Then the corresponding

$$\min_\lambda g(\lambda) = b^T \lambda$$

where $A^T\lambda = c$ and $\lambda \geq 0$ is what we call the **dual problem**(D)

Theorem 17 (Strong Linear Programming Duality)
If (P) admits an optimal solution x^* then (D) admits an optimal solution λ^* such that

- the values of the two problems coincide that is

$$g(\lambda^*) = (\lambda^*)^T b = c^T x^* = f(x^*)$$

- the solution λ^* satisfies complementary slackness condition that is

$$(\lambda^*)^T(b - Ax^*) = 0$$

Proof. From first order *necessary* optimality conditions for (P) we know that any optimal x^* must satisfy the condition

$$\nabla f(x^*) \in \mathcal{N}_\Omega(x)$$

where $\Omega = \{x \in \mathbb{R}^m : Ax \leq b\}$. From 31 and 32 we have

$$\mathcal{N}_\Omega(x^*) = \left\{A^T\lambda : \lambda^T(b - Ax^*) = 0, \lambda \geq 0\right\}$$

combining these 2 equations and using the fact that $\nabla f(x^*) = c$ we obtain that at optimality there must exist $\lambda^* \in \mathbb{R}^n$ such that

$$c = A^T\lambda^*, (\lambda^*)^T(b - Ax^*), \lambda^* \geq 0$$

but this may be simplified further to

$$g(\lambda^*) = (\lambda^*)^T b = c^T x^* = f(x^*)$$

It now suffices to show that λ^* is indeed the optimal value for (D) which follows from

$$\begin{aligned} g(\lambda) &= b^T\lambda \\ &\geq (Ax^*)^T\lambda \\ &= (x^*)^T A^T \lambda \\ &= (x^*)^T c = f(x^*) = g(\lambda^*) \end{aligned}$$

now because the choice of λ was arbitrary we have our conclusion as desired □

3 The special case of convex functions(3)

3.1 convex functions

Recall from above that any solution x to a nonlinear optimization problem defined on a convex feasible set $\Omega \subseteq \mathbb{R}^n$ must necessarily satisfy the first order optimlaity condition

$$\langle \nabla f(x), y - x \rangle \geq 0 \quad \forall y \in \Omega$$

In general this condition is only *necessary* but *not suffcient*. However there exists a notable class of functions for which such a condition is *sufficient*. These are called **convex functions** and will be the focus of our discussion for this section.

Definition 18

Let $\Omega \subseteq \mathbb{R}^n$ be convex. A function $f : \Omega \to \mathbb{R}$ is **convex** if for any two points $x, y \in \Omega$ and $t \in [0, 1]$,

$$f((1-t) \cdot x + t \cdot y) \leq (1-t) \cdot f(x) + t \cdot f(y)$$

Example 19

Consider the following example of a convex function

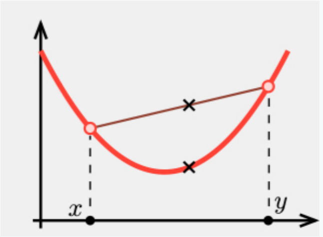


Figure 1: Convex function

You could say the sum of the function values at the end points will always be greater than the function value at any point on the segment in between these end points

Example 20

Consider a few more examples of convex functions for further visualization

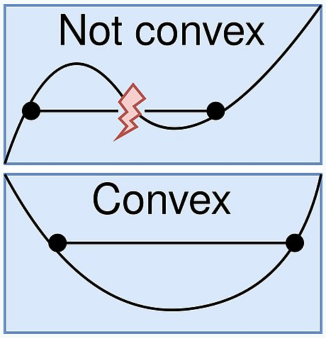


Figure 2: Convex vs non convex

You could say that for a function to convex if and only if the line segment connecting any two points on the graph lies above or on the graph. So basically a convex function is one where there are only "flat ground" or "troughs" and no "hills"

These observations has another important implication. That is

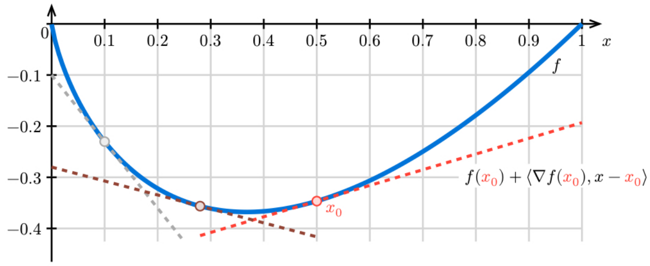


Figure 3: Convexity implies bounding by linearization

Formally we state

Theorem 21

Let $f : \Omega \to \mathbb{R}$ be convex and differentiable function defined on a convex domain Ω. Then for all $x \in \Omega$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall y \in \Omega$$

Proof. Pick any $x, y \in \Omega$. By definitin of convexity we have

$$f(x + t \cdot (y - x)) \leq f(x) + t \cdot (f(y) - f(x)) \quad \forall t \in [0, 1]$$

Then moving $f(x)$ from the RHS to the LHS then dividing by t we have

$$\frac{f(x + t \cdot (y - x)) - f(x)}{t} \leq f(y) - f(x) \quad \forall t \in (0, 1]$$

Then taking a limit as $t \downarrow 0$ and recoginizing a directional derivatve at x along direction $y - x$ on the LHS we conclude that

$$\langle \nabla f(x), y - x \rangle \leq f(y) - f(x)$$

Rearranging yields the result

Theorem 22

Let $\Omega \subseteq \mathbb{R}^n$ be convex and $f : \Omega \to \mathbb{R}$ be a convex differentiable function. Then $-\nabla f(x) \in \mathcal{N}_\Omega(x) \quad \Leftrightarrow$ x is a minimizer of f on Ω

Proof. We already know from previou sections that $-\nabla f(x) \in \mathcal{N}_\Omega(x)$ is necessary for optimality. So we just need to show sufficiency. Specifically we need to show that if $\langle f(x), y - x\rangle \geq 0$ for all $y \in \Omega$ then surely $f(y) \geq f(x)$ for all $y \in \Omega$. However that is precisely what we have shown in 21 so the theorem follows.

3.2 equivalent defintions of convexity

Fact 23 (General Inequalities)

Notation matters:

$A \succ 0$: This denotes that A is **positive definite** (PD). This means $z^T Az > 0$ for all $z \neq 0$, implying that A has strictly positive eigenvalues.

$A \succeq 0$: This denotes that A is **positive semidefinite** (PSD), meaning that $z^T Az \geq 0$ for all z, implying that A has non-negative eigenvalues.

$A \preceq 0$: This denotes that A is **negative semidefinite** (NSD), meaning that $z^T Az \leq 0$ for all z.

$A \prec 0$: This denotes that A is **negative definite** (ND), meaning that $z^T Az < 0$ for all $z \neq 0$.

Theorem 24

Let $\Omega \subseteq \mathbb{R}^n$ be a convex set and $f : \Omega \to \mathbb{R}$ be a function. The following are equivalent definitions of convexity

1. $f((1-t)\cdot x + t \cdot y) \leq (1-t)\cdot f(x) + t \cdot f(y)$ for all $x, y \in \Omega, t \in [0, 1]$
2. (If f is differentiable) $f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle$ for all $x, y \in \Omega$
3. (If f is twice differentiable and Ω is open) $\nabla^2 f(x) \succcurlyeq 0$ for all $x \in \Omega$

Remark 25. *In practice, (1) is the most general condition while (2) is the widely used condition while (3) is the easiest is the easiest to check*

Remark 26. *Note that $\nabla^2 f$ here refers to the* **hessian** *which is the gradient applied twice not the Laplacian. So in our case the hessian of some scalar function is therefore some matrix consisting of all mixed order 2 partial differentials $\frac{\partial f}{\partial x_i \partial x_j}$. That is why in practice we write $\Delta = \nabla \cdot \nabla f(x)$(divergence of gradient) to refer to the laplacian to avoid confusion.*

Proof. We have already seen how (1) $\Rightarrow$ (2) in 22. So we now show that under differentiability (2) $\Rightarrow$ (1) and that under twice differentiability and openness of Ω, we have (2) $\Leftrightarrow$ (3). We break the proof into separate steps

- *Proof.* that (2) $\Rightarrow$ (1). Pick any $a, b \in \Omega$ and $t \in (0, 1)$ and consider the point

$$z = t \cdot a + (1-t)\cdot b \in \Omega$$

This exists by the definition of convex set. Therefore since $a, b, z \in \Omega$ we have from the linearization bound (2) for the choices $(x, y) = (z, a), (z, b)$ we know that

$$f(a) \geq f(z) + \langle \nabla f(z), a - z \rangle$$
$$f(b) \geq f(z) + \langle \nabla f(z), b - z \rangle$$

Multiplying the first inequality by t and the second by $1 - t$ and summing we obtain

$$t \cdot f(a) + (1 - t) \cdot f(b) \geq f(z) + \langle \nabla f(z), \underbrace{t \cdot a + (1 - t) \cdot b - z}_{=0} \rangle = f(z)$$

then clearly (1) follows

Proof. that (2) ⇒ (3). Pick any two $x, y \in \Omega$ and $t \in (0, 1]$. Define $x_t = x + t \cdot (y - x)$ where $x_t \in \Omega$ by definition of convexity of Ω just like before. From (2) we can write

$$f(x_t) \geq f(x) + \langle \nabla f(x), x_t - x \rangle$$
$$f(x) \geq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$$

By bilearity of the inner product we can also write

$$f(x) \geq f(x_t) + \langle -\nabla f(x - t), x_t - x \rangle$$

then summing the inequalities we have

$$\begin{aligned} 0 &\geq \langle \nabla f(x) - \nabla f(x_t), x_t - x \rangle \\ &= \langle \nabla f(x) - \nabla f(x_t), t \cdot (y - x) \rangle \\ &= t \cdot \langle \nabla f(x) - \nabla f(x_t), y - x \rangle \end{aligned}$$

Dividing both sides by t^2 we have

$$\frac{\langle \nabla f(x + t \cdot (y - x)) - \nabla f(x), y - x \rangle}{t}$$

then multiplying the 1st argument of the inner product by -1(and changing the sign of the inquality accordingly) and taking the limit as $t \downarrow 0$ we therefore have

$$\langle (y - x), \nabla^2 f(x)(y - x) \rangle \geq 0$$

Because Ω is open, that means that every x has a neighborhood contained in Ω(i.e x is an interior point). And because the choice of y arbitrary we are able to choose points y all around the point x and therefore the direction $y - x$ is abirtrary as well. This the hessian matrix $\nabla^2 f(x)$ is **positive semi-definite**

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

as desired

Proof. that (3) ⇒ (2) Well by observation we know some kind of double integration is probably involved. Now let's again pick any $x, y \in \Omega$ and $\tau \in [0, 1]$. Now since $x + \tau \cdot (y - x)$ must also be in Ω by definition of convexity, by (2) we know $\nabla^2(x+\tau\cdot(y-x)) \succcurlyeq 0$ so we have

$$0 \leq \langle y-x, \nabla^2 f(x+\tau\cdot(y-x))\cdot(y-x)\rangle = (y-x)^T \underbrace{\nabla^2 f(x+\tau\cdot(y-x))(y-x)}$$

Note that the expressions in the bracket is equal to the dot product since $\nabla^2 f(x+\tau\cdot(y-x)) = (\nabla^2 f(x+\tau\cdot(y-x)))^T$. This is because f is C^2 and therefore the order of mixed partial derivatives are immaterial refer to your `18.101 Analysis 2` notes for proof. Hence taking the double integral we have

$$\begin{aligned}
0 &\geq \int_0^1 \int_0^t \langle y-x, \nabla^2 f(x+\tau\cdot(y-x))\cdot(y-x)\rangle d\tau dt \\
&= \int_0^1 \langle y-x, \int_0^t \underbrace{\nabla^2 f(x+\tau\cdot(y-x))\cdot(y-x)}_{\frac{d}{d\tau}\nabla f(x+\tau\cdot(y-x))} d\tau\rangle dt \\
&= \int_0^1 \langle y-x, \nabla f(x+t\cdot(y-x)) - \nabla f(x)\rangle dt \\
&= -\langle \nabla f(x), y-x\rangle + \int_0^1 \underbrace{\langle \nabla f(x+t\cdot(y-x)), y-x\rangle}_{\frac{d}{dt} f x+t\cdot(y-x))} dt \\
&= f(y) - f(x) - \langle \nabla f(x), y-x\rangle
\end{aligned}$$

upon rearrangement obtains (2)

□

In addition to the above criteria we also have **convex preserving operations**

Theorem 27

The following operations preserve convexity

- Mulitplication of a convex function $f(x)$ by a nonnegative scalar $c \geq 0$
- Addition of two convex function $f(x), g(x)$
- pointwise suprenum of a collection J of convex functions $\{f_j(x) : j \in J\}$

$$f_{\max}(x) = \max_{j\in J} f_j(x)$$

- Pre-composition $f(Ax+b)$ of a function with an affine function $Ax+b$
- Post-composition $g(f(x))$ of a convex function with an increasing function g

Proof. All these can be readily verified from the definition of a convex function

4 Feasibility,optimization and separation

So far we have introduced the study of non linear optmization problems as well as explored various first order optimality conditions. We now show that in many cases(not all) the solution to a convex optimization problem

$$\min_x f(x) \text{ convex}$$

where $x \in \Omega$ convex can be found in **polynomial** time

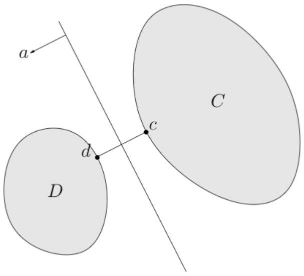
4.1 Appendix: Separating Hyperplane theorem

Theorem 28

Suppose C and D are nonempty disjoint convex sets i.e ($C \cap D = \emptyset$) Then there exists $a \neq 0$ and b such that $a^T b \leq b$ for all $x \in C$ and $a^T x \geq b$ for all $x \in D$.The hyperplane $\left\{x | a^T x = b\right\}$ is called the **separating hyperplane** for the sets C and D

Proof. First we assume the the (eucidlean) distance between C and D is defined as

$$\textbf{dist}(C, D) = \inf \left\{ ||u - v||_2 \, | u \in C, v \in D \right\}$$



Now define...to be continued just refer to the stanford convex analysis course book by prof Stephen Bloyd or convince yourself intuitively using the below.

4.2 Separating a point from a closed convex set

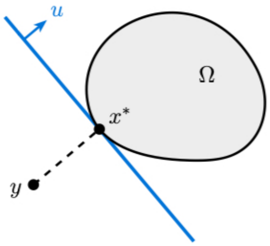
Consider a special case of the separating hyperplane theorem

Theorem 29

Let $\Omega \subseteq \mathbb{R}^n$ be a nonempty,closed and convex set and let $y \in \mathbb{R}^n$ be a point. If $y \neq \Omega$ then there exist $u \in \mathbb{R}^n, v \in \mathbb{R}$ such that

$$\langle u, y \rangle < v \quad \text{and} \quad \langle u, x \rangle \geq v \quad \forall x \in \Omega$$

Proof. Consider a point x^* on the boundary. Then define the halfspace such that is is orthogonal to the line that joins x^* and y(outside Ω). Now we know that from 2 there exists a projection point $x^* \in \Omega$ such that it is the solution to $\min_x \frac{1}{2} ||x - y||_2^2$. The gradient of this objective function is as seen previously $x - y$.



so by first order conditions

$$\langle x^* - y, x - x^* \rangle \geq 0 \quad \forall x \in \Omega$$

Let now $u = x^* - y$ and $v = \langle u, x^* \rangle$. This implies

$$\langle u, x \rangle - \langle u, x^* \rangle \geq 0 \quad \Rightarrow \quad \langle u, x \rangle \geq v$$

Also we have

$$\langle u, y \rangle = \langle u, x^* - u \rangle = v - ||u||_2^2 < v$$

□

This theorem justifies the following definition.

Definition 30 (Strong separation oracle)
Let $\Omega \subseteq \mathbb{R}^n$ be convex and compact. A **strong separation oracle** for Ω is an algorithm that given any point $y \in \mathbb{R}^n$ correctly outputs one of the following

- $y \in \Omega$
- $(y \notin \Omega, u, v)$ where the pair $(u, v) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$\langle u, y \rangle < \quad \text{and} \quad \langle u, x \rangle \geq v \quad \forall x \in \Omega$$

...to be continued

5 Lagrange multipliers and KKT conditions(5)

5.1 a second look at the normal cone of linear constraints

Recall that in lecture (2) we considered normal cones for a few classes of feasible sets that come up often like hyperplanes,affine subspsaces,halfspaces and intersection of halfspaces

Theorem 31
Let $\Omega \subseteq \mathbb{R}^n$ be defined as the intersection of m linear inequalities

$$\Omega = \{x \in \mathbb{R}^n : Ax \leq b\} \text{ where } A = \begin{pmatrix} -a_1^T- \\ \vdots \\ -a_m^T- \end{pmatrix} \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$$

Given a point $x \in \Omega$ define the index set of the binding constraints(border lines) defined by

$$I(x) = \left\{j \in \{1, \dots, m\} : a_j^T x = b_j\right\}$$

Then the normal cone at any $x \in \Omega$ is given by

$$\mathcal{N}_\Omega(x) = \left\{ \sum_{j \in I(x)} \lambda_j a_j : \lambda_j \geq 0 \right\} = \left\{A^T \lambda : \lambda^T(b - Ax) = 0, \lambda \in \mathbb{R}^m_{\geq 0}\right\}$$

Definition 32

In the definition of $\mathcal{N}_\Omega$ above instead of summing over $j \in I(x)$ one could equivalently sum over al $j = 1, \ldots, m$ and then impose the constraint that $\lambda_j = 0$ for all $j \neq I(x)$. That is

$$\mathcal{N}_\omega(x) = \left\{ \sum_{j=1}^{m} \lambda_j \cdot a_j : \lambda_j \geq 0, \lambda_j = 0 \text{ if } \langle a_j, x \rangle < b_j \right\}$$

Because $\lambda_j \geq 0$ the condition that $\lambda_j = 0$ whenever $\langle a_j, x \rangle < b_j$ can be succintly written as

$$\sum_{j=1}^{m} \lambda_j (b_j - \langle a_j, x \rangle) = 0$$

the above condition is usually called **complementary slackness**

Consider an example application of the above theorem 31

Example 33

Consider the intersection Ω of two halfspaces

$$\Omega = \left\{ y \in \mathbb{R}^n : \begin{array}{c} \langle a_1, y \rangle \leq b_1 \\ \langle a_2, y \rangle \leq b_2 \end{array} \right\}$$

where $a_1, a_2, \in \mathbb{R}^n$ and $a_1, a_2 \neq 0$(essentially they are the normals of half planes)

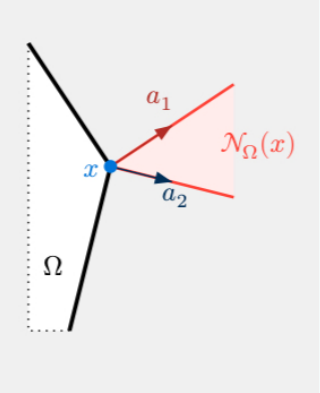


Figure 4: intersection halfspace

Let x be a point in Ω.If x is in the interior of Ω or if x is on the boundary of one of the halfspaces but not both then the normal cone follows from our prior results from earlier lectures. So we consider x at the intersection of both halfspaces that is $\langle a_1, x \rangle = b_1, \langle a_2, x \rangle = b_2$(since they are at the boundary). Now on application of 31 we get

$$\mathcal{N}_\Omega = \{\lambda_1 \cdot a_1 + \lambda_2 \cdot a_2 : \lambda_1, \lambda_2 \in \mathbb{R}_{\geq 0}\}$$

which visually makes sense too. $\mathcal{N}_\Omega(x)$ here is what we call the **conic hull** of a_1 and a_2

We will prove this theorem later below

Definition 34 (Cone)

A set S is a **cone** if any $x \in S$ and $\lambda \in \mathbb{R}_{\geq 0}$ the point $\lambda \cdot x \in S$

Theorem 35

Let $S \subseteq \mathbb{R}^n$ be a nonempty closed convex cone and $y \notin S$ be a point in $\mathbb{R}^n$. Then there exists a hyperplane passing through the origin that separates y from S formally there exists $u \in \mathbb{R}^n$ such that

$$\langle u, y \rangle < 0 \quad \text{and} \quad \langle u, x \rangle \geq 0 \quad \forall x \in S$$

Proof. We already know from 29 that there exists $u \in \mathbb{R}^n, v \in \mathbb{R}$ such that

$$\langle u, v \rangle < v \quad \text{and} \quad \langle u, x \rangle \geq v \quad \forall x \in S$$

Now consider that the separation condition on the right of the above implies that $v \leq \lambda \cdot \langle u, a \rangle$ for all $\lambda \geq 0$. In particular when $\lambda = 0$ we see that $v \leq 0$ yielding $\langle u, v \rangle < 0$. Furthermore dividing by λ we find that

$$\langle u, a \rangle \geq \frac{v}{\lambda} \quad \forall \lambda \geq 0 \quad \Rightarrow \quad \langle u, a \rangle \geq \sup_{\lambda \to \infty} \frac{v}{\lambda} = 0$$

Since $a \in S$ is arbitary the statement follows □

Now we are ready to prove theorem 31.

Proof. Fix any $x \in \Omega$ and let

$$\mathcal{C}(x) = \left\{ \sum_{j \in I(x)} \lambda_j a_j : \lambda_j \geq 0 \right\}$$

We will now show that $\mathcal{N}_\Omega(x) = \mathcal{C}(x)$ by proving the 2 directions of inclusion.

- Showing that $d \in \mathcal{C}(x)$ belongs to $\mathcal{N}_\Omega(x)$ that is

$$\langle d, y - x \rangle \leq 0 \quad \forall y \in \Omega$$

  Let d be expressed as $\sum_{j \in I(x)} \lambda_j a_j$ with $\lambda_j \geq 0$ then for an $y \in \Omega$

$$\begin{aligned} \langle \sum_{j \in I(x)} \lambda_j a_j, y - x \rangle &= \sum_{j \in I(x)} \lambda_j \langle a_j, y - x \rangle \\ &= \sum_{j \in I(x)} \lambda_j (\langle a_j, y \rangle - b_j) \\ &\leq \sum_{j \in I(x)} \lambda_j (b_j - b_j) = 0 \end{aligned}$$

  where the second line follows by definition of $I(x), \langle a_j, x \rangle = b_j$ and the final line follows since $y \in \Omega$ and $\lambda_j \geq 0$. This shows that $d \in \mathcal{N}_\Omega(x)$

- from the other direction. We aim to show the contrapositive $d \notin \mathcal{C}(x) \Rightarrow d \notin \mathcal{N}_\Omega(x)$. Assuming such a d since $\mathcal{C}$ is a nonempty closed convex cone, by the conic separatiuon result in the previous theorem there exists $u \in \mathbb{R}^n$ such that

$$\langle u, d \rangle < 0 \quad \text{and} \quad \langle u, a \rangle \geq 0 \quad \forall a \in \mathcal{C}(x)$$

We argue that for $\delta > 0$ small enough the point $y = x - \delta \cdot u$ belongs to Ω. That is δ satisfies for any given $x \in \Omega$(considering all cases whether or not $j \in I(x)$)

1. if $j \in I(x)$ then $\langle a_j, x - \delta \cdot u \rangle = b_j - \delta \cdot \langle a_j, u \rangle \leq b_j$ since $\langle a_j, u \rangle \geq 0$ by the above. So we now have precisely the criterion to remain in Ω
2. if $j \notin I(x)$ then $b_j - \langle a_j, x \rangle > 0$(since $x \in \Omega \Rightarrow \langle a_j, x \rangle \leq b_j$ and also note that a_i is a particular vector of 33 so it is not a coordinate of a but rather a particular instance of $a \in C(x)$). This means looking at $\langle a_j, x - \delta \cdot u \rangle \leq b_j - \delta \cdot \langle a_j, u \rangle \leq b_j$ no matter the value of δ , $y \in \Omega$

Now that means

$$\langle d, y - x \rangle = \langle d, -\delta \cdot u \rangle = -\delta \cdot \langle d, u \rangle > 0$$

where both $y, x \in \Omega$ since δ small enough. Notice we now have precisely the criterion that says $d \notin \mathcal{N}_\Omega(x)$ as desired.

5.2 Karush-Kuhn-Tucker(KKT) conditions

The result of 31 gives a complete characterization of the normal cone for sets defined as the intersections of linear contraints. We now turn our attention to more general constaint sets defined as the intersection of differentiable functional constraints

$$\min_x f(x)$$

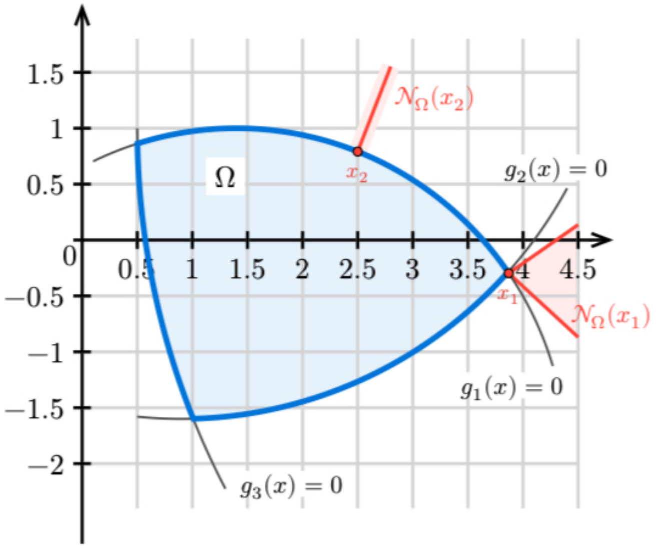
where

$$h_i(x) = 0 \quad i \in \{1, \ldots, r\}$$

and

$$g_j(x) = 0 \quad j \in \{1, \ldots, s\}$$

The General Idea



Consider any point x^* on the boundary of the feasible set Ω depicted above which is the intersection of three inequality constraints $g_i(x) \leq 0$ for $i = 1, 2, 3$.

Fact 36 (Key Observation)
If you had not noticed by now, observe that the set of directions that form an obtuse angle with all directions that from x^* remain inside of the set coincides with the normal cone of the linearization of the constraints that are binding(that is holding with equality) at x^*

This is why we used δ small enough such that $y \in \Omega$ in our proof of theorem 31.

5.3 failure of KKT conditions

However in general, KKT conditions might fail to hold at optimality. For the rest of this section, this will be 2nd part of an appendix I first created in my `Stanford Convex Analysis` notes(material is sourced from `James V. Burke`'s Nonlinear Optimization UoW course notes). To recap, in part 1 we have covered the following key points

1. The logical proposition $P \rightarrow Q$ means that
   - Q is a **necessary condition** of P
   - P is **sufficient condition** for Q
2. KKT conditions are only necessary conditions for optimality under certain what we call "constraint qualifications/regulatory conditions" - we saw an example of how failure of LICQ resulted in failure of KKT and proven how regularity made KKT a necessary condition for optimality
3. In `Stanford Convex Analysis` we have covered Slater's condition for convex optimization problems rigorously. Today we aim to cover the other widely applicable contraints that we spoke of namely:
   (a) Linear Contraint qualification(LCQ)
   (b) Linear Independence contraint qualification(LICQ)
   (c) Mangasarian-Fromovitz constraint qualification(MFCQ)

Before we proceed i have copied over some definitions from part 1 for ease of reference here

Definition 37 (Feasible directions)
Given a subset $\Omega \in \mathbb{R}^n$ and a point $x \in \Omega$ we say that a direction $d \in \mathbb{R}^n$ is a **feasible direction** for Ω at x if there is a $\bar{t} > 0$ such that $x + td \in \Omega$ for all $t \in [0, \bar{t}]$

First define the nonlinear optimization problem(NLP)

$$\min f(x)$$

Definition 38 (Feasible Set of NLP)
Our feasible set Ω is such that every $x \in \Omega$ satisfies the contraints

$$\begin{aligned} c_i(x) &\leq 0 \qquad i = 1, \ldots, s \\ c_i(x) &= 0 \quad i = s+1, \ldots, m \end{aligned}$$

where $c_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable on $\mathbb{R}^n$.

Definition 39

The **tangent cone** to Ω at a point $x \in \partial\Omega$(boundary of feasible set) is the set of limiting directions obtained from sequences in Ω that converge to x. That is

$$T(x|\Omega) = \left\{d : \exists \tau_i \searrow 0 \text{ and } \{x_i\} \subset \partial\Omega \text{ with } x_i \to x \text{ such that } \tau_i^{-1}(x_i - x) \to d\right\}$$

Notice that $T(x|\Omega)$ is clearly a closed cone since all its limit points are defined to be in it.

Example 40

Consider

1. if $\Omega = \{x \in \mathbb{R}^n | \, ||x||_2 = 1\}$ then $T(x|\Omega) = \{v \in \mathbb{R}^n | \langle v, x\rangle = 0\}$.In particular when $n = 2$ then $\Omega = \left\{(x_1, x_2)^T | x_1^2 + x_2^2 = 1\right\}$ and $T(x|\Omega) = \left\{v \in \mathbb{R}^2 | v_1 x_1 + v_2 x_2 = 0\right\}$ is the line tangent to Ω at $(x_1, x_2)^T$ translated to pass through the origin(can translate since we just considering directions not actual points)
2. for a general manifold the tangent cone is simply the tangent space

Definition 41

Let $d \in T(x|\Omega)$.Let a sequence $\{x_k\} \subset \Omega$ and $\tau_k \searrow 0$ with $x_k \to x$ such that ${\tau_k}^{-1}(x_k - x) \to d$. Then setting $d_k = {\tau_k}^{-1}(x_k - x)$ for all k we have that

$$c_i'(x; d) = \lim_{k \to \infty} \frac{c_i(x + \tau_k d_k)}{\tau_k}$$

Definition 42

we denote

$$I(x) = \{i : i \in \{1, \ldots, s\}, c_i(x) = 0\}$$

for each x. The purpose of doing so is to only consider active constraints for convenience and efficiency. Because it is inequality, the contraint is not "active" in a sense you are not yet on the boundary which is at 0

Definition 43

consequently defining the **linearized cone**

$$L_\Omega = \left\{d : \nabla c_i(x)^T d \leq 0, i \in I(x), \nabla c_i(x)^T d = 0, i = s+1, \ldots, m\right\}$$

basically it means if you take a small positive step in direction d and you still satisfy 1st order(linear) constraints. Unlike $T(x|\Omega)$(the tangent cone) it does not represent all feasible directions. For example it does not consider curvature or geometry since it is but a mere linearization.(note that if you take a linear step on the boundary you can never go into the interior by definition tangent)

we have that in general

$$\boxed{T(x|\Omega) \subset L_\Omega(x)}$$

To see this consider

Example 44

note that (a) is convex (b) is not

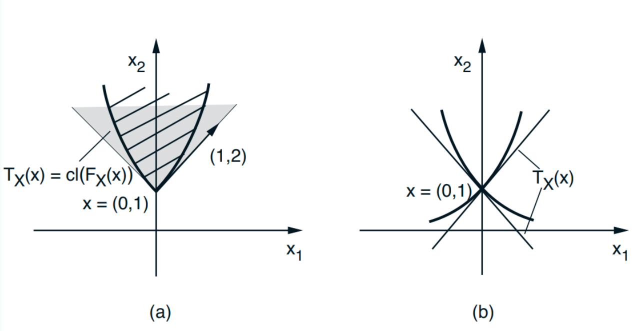


Figure 5: in here $T_X(x)$ refers to the 1st order feasible directions

If you recall this satisfies slater condition since

- convex optimization problem
- feasible set has an interior point

so thats why (a) is regular(see below for defintion)

Definition 45 (Regularity)

We say the representation of the set Ω is **regular** at $x \in \Omega$ if

$$T(x|\Omega) = \{d \in \mathbb{R}^n : c_i'(x;d) \le 0, i \in I(x), c_i'(x;d) = 0 \ i = s+1, \dots, m\}$$

In other words $T(x|\Omega) = L_\Omega(x)$

Example 46

Let $\Omega = \left\{x \in \mathbb{R}^2 | (x_1^2 + x_2^2) - 1 \le 0, ((x_1 - 2)^2 + x_2^2) - 1 \le 0\right\}$ so that Ω is the single point $\bar{x} = (1,0)^T$. In this case $T\{\bar{x}|\Omega\} = \{0,0\}^T$ while $L_\Omega(\bar{x}) = \left\{(0,\lambda)^T | \lambda \in \mathbb{R}\right\}$. However if we simply represent Ω as $\Omega = \{x | x_1 = 2, x_2, 0\}$ then $T(x|\Omega) = L\Omega(x)$

$(x_1^2 + x_2^2) - 1 \le 0$: This defines a disk of radius 1 centered at $(0,0)$.

$((x_1 - 2)^2 + x_2^2) - 1 \le 0$: This defines a disk of radius 1 centered at $(2,0)$.

The feasible set Ω is the intersection of these two disks. Since the disks are tangent at the point $(1,0)$, the feasible set Ω consists of only this single point:

$$\Omega = \{\bar{x} = (1,0)^T\}.$$

The tangent cone $T(x|\Omega)$ captures the directions of feasible "motion" near $\bar{x}$. Since Ω is just the single point $(1,0)^T$, there is no direction in which we can move within Ω while remaining feasible. Therefore, the tangent cone is trivial:

$$T(\bar{x}|\Omega) = \{(0,0)^T\}.$$

Now consider the linearization $L_\Omega(x)$, which is defined by the gradient conditions:

$$L_\Omega(\overline{x}) = \{d : \nabla c_i(\overline{x})^T d \leq 0 \text{ for } i \in I(\overline{x}), \nabla c_i(\overline{x})^T d = 0 \text{ for } i = s+1, \dots, m\}.$$

At $\overline{x} = (1,0)^T$, compute the gradients of the two constraints: For the first constraint, $(x_1^2 + x_2^2) - 1 = 0$ at $(1,0)$:

$$\nabla c_1(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \Big|_{(1,0)} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

For the second constraint, $((x_1 - 2)^2 + x_2^2) - 1 = 0$ at $(1,0)$:

$$\nabla c_2(x) = \begin{bmatrix} 2(x_1 - 2) \\ 2x_2 \end{bmatrix} \Big|_{(1,0)} = \begin{bmatrix} -2 \\ 0 \end{bmatrix}.$$

These gradients $\nabla c_1(x)$ and $\nabla c_2(x)$ are collinear but opposite in direction. The linearized feasible directions are determined by satisfying:

$$\nabla c_1(x)^T d \leq 0 \quad \text{and} \quad \nabla c_2(x)^T d \leq 0.$$

This implies:

$$2d_1 \leq 0 \quad \text{and} \quad -2d_1 \leq 0.$$

The only solution is $d_1 = 0$, leaving d_2 unconstrained. Thus:

$$L_\Omega(\overline{x}) = \{(0, \lambda)^T : \lambda \in \mathbb{R}\}.$$

Clearly, $T(\overline{x}|\Omega) = \{(0,0)^T\}$ is strictly smaller than $L_\Omega(\overline{x}) = \{(0,\lambda)^T : \lambda \in \mathbb{R}\}$. This discrepancy arises because the linearization L_Ω fails to capture the geometric reality that Ω is a single point. Hence, the representation of Ω is not regular at $\overline{x}$.

Now Analyze Ω under the simpler representation Now, represent Ω as:

$$\Omega = \{x : x_1 = 1, x_2 = 0\}.$$

For this representation, the constraints are explicit:

$$x_1 - 1 = 0, \quad x_2 = 0.$$

$c_1(x)$ constrains x_1 to 1. $c_2(x)$ constrains x_2 to 0. Thus, Ω is precisely the point $(1,0)$.

Since Ω is just the single point $(1,0)$, there are no other points in Ω that can form a sequence approaching $\overline{x}$. Therefore, the only "direction" available is $(0,0)$:

$$T(\overline{x}|\Omega) = \{(0,0)^T\}.$$

The linearization $L_\Omega(\overline{x})$ is determined by the gradients of the active constraints at $\overline{x}$:

$$L_\Omega(\overline{x}) = \{d : \nabla c_i(\overline{x})^T d \leq 0,\ i \in I(\overline{x});\ \nabla c_i(\overline{x})^T d = 0,\ i = s+1, \dots, m\}.$$

Here, all constraints are equality constraints ($c_1(x) = 0$ and $c_2(x) = 0$), so we solve:

$$\nabla c_1(\overline{x})^T d = 0, \quad \nabla c_2(\overline{x})^T d = 0.$$

For $c_1(x) = x_1 - 1$:

$$\nabla c_1(x) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

For $c_2(x) = x_2$:

$$\nabla c_2(x) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

To satisfy $\nabla c_1(\overline{x})^T d = 0$ and $\nabla c_2(\overline{x})^T d = 0$, we solve:

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = 0 \quad \text{and} \quad \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \end{bmatrix} = 0.$$

This implies:

$$d_1 = 0, \quad d_2 = 0.$$

Thus:

$$L_\Omega(\overline{x}) = \{(0,0)^T\}.$$

Since Ω is defined as the single point $(1,0)$, the tangent cone $T(\overline{x}|\Omega)$ and the linearized feasible directions $L_\Omega(\overline{x})$ both correctly represent the trivial geometry of Ω:

$$T(\overline{x}|\Omega) = L_\Omega(\overline{x}) = \{(0,0)^T\}.$$

So Ω has regular representation

Example 47

Recall an example that we have covered earlier in 33. We now denote that the green lines represent

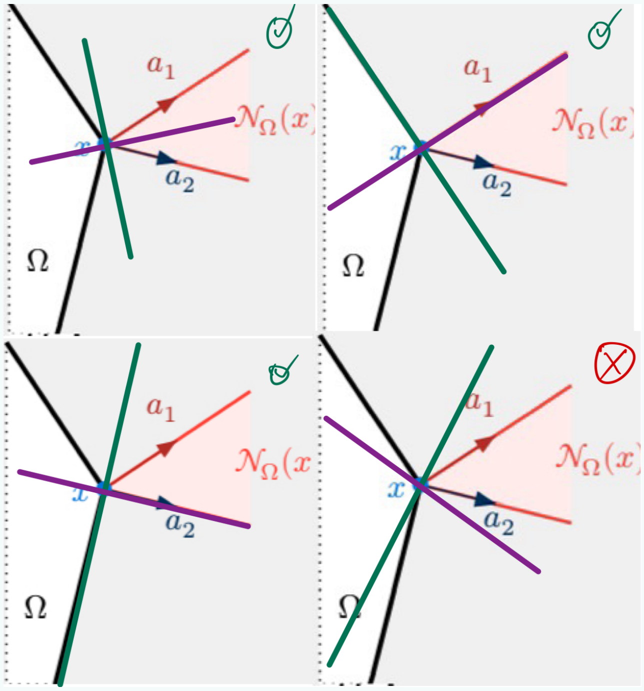


Figure 6: Notice that the red zone is the normal cone. And the grey zone is the tangent cone. The set of feasible directions either lie on the boundary of Ω or in the grey zone. Never insider Ω or $\mathcal{N}_\Omega$

We were able to solve for KKT/langrange multipliers because Ω was regular. To see this consider that the linearization

$$L_\Omega = \left\{d : a_1^T d \leq 0, a_2^T d \leq 0\right\}$$

however that is precisely the tangent cone which is defined in this case

5.4 LICQ and MFCQ

Definition 48 (LICQ and MFCQ Contraint Qualifications)
Let $\overline{x}$ be feasible for our (NLP) and put $I(\overline{x}) = \{i | c_i(\overline{x}) = 0, i = 1, 2, \ldots, s\}$. We say that

1. the **linear independence constraint qualification**(LICQ) holds at $\overline{x}$ if the gradients

$$\nabla c_i(\overline{x})(i \in I(\overline{x})) \quad \nabla c_j(\overline{x})(i = s+1, \ldots, m)$$

are linearly independent

2. The **Mangasarian Fromovitz contraint qualification**(MFCQ) holds at $\overline{x}$ if the gradients

$$\nabla c_i(\overline{x})(i = s+1, \ldots, m)$$

are linearly independent and there exists a vector $d \in \mathbb{R}^n$ such that

$$\nabla c_i(\overline{x})^T d < 0(i \in I(\overline{x}) \quad \nabla c_i(\overline{x})^T d = 0(j = s+1, \ldots, m) \tag{1}$$

Lemma 49 (The Access Lemma for NLP)
Let $\overline{x} \in \Omega$ be such that MFCQ is satisfied at $\overline{x}$. Then for every direction of d satisfying 1 there exists $\varepsilon > 0$ and a C^1 curve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $x(t) \in \Omega$ for all $t \in [0, \varepsilon), x(0) = \overline{x}$ and $x'(0) = d$

Proof. Define $C_e : \mathbb{R}^n \to \mathbb{R}^{(m-s)} \to \mathbb{R}^{(m-s)}$ by

$$C_e(x) := (c_{s+1}(x), \ldots, c_m(x))^T$$

and define $\hat{C}_i(y, t) := \mathbb{R}^{(m-s)+1} \to \mathbb{R}^{(m-s)}$ by

$$\hat{C}_i(y, t) = c_i(\overline{x} + td + \nabla C_e(\overline{x})^T y) \quad i = s+1, \ldots, m$$

where $\nabla C_e(\overline{x})$ denotes the Jacobian of C_e at $\overline{x}$.

Remark 50. *You know that by convention it must take the same shape as $C_e(x)$(that is it must have m-s rows). Now is a good time to quickly review your* `MIT 18.S096(2023) Matrix Calculus` *notes if you cant remember this.*

The nonlinear equation $\hat{C}(y, t) = 0$ has the solution $(\overline{y}, \overline{t}) = (0, 0)$(recall $\overline{x} \in \Omega$ means to satisfy contraints it must have $c_i(\overline{x}) = 0$) with

$$\nabla_y \hat{C}(0, 0) = \nabla C_e(\overline{x}) \nabla C_e(\overline{x})^T$$

To see this note that

- $\nabla_y \hat{C}_i(0, 0) = \nabla_y c_i(\hat{x}) \nabla C_e(\overline{x})$
  by chain rule and subbing in $(y, t) = (0, 0)$. Now letting i range across all possible i(replace it with ":") we can express the matrix

$$\nabla_y \hat{C}(0, 0) = \nabla C_e(\overline{x}) \nabla C_e(\overline{x})^T$$

  as above

- $\nabla_t \hat{C}_i(0,) = c_i(\hat{x}) d$
  Once again by chain rule. Now similar to above we then have

$$\nabla C_e(\overline{x}) d$$

but by assumption that $\nabla c_i(\overline{x})(i = s{+}1, \dots, m)$ are linearly independent we know that this matrix is non-singular.Then the **implicit function theorem**

Remark 51. *You should know that a matrix product of invertible matrix is invertible. Just see that* $\det(AB) = \det A \det B \neq 0$. *And you also know that the tranpose of an invertible matrix is invertible. Just see that* $\det A = \det A^T$

Remark 52. *implicit function theorem was derived in your `MIT 18.101 Analysis 2` notes. I know what you are thinking but don't get too excited. Though it is not to be confused with inverse function theorem, it is used to derive implict function theorem. You cannot escape the machinery lmao*

yields a C^1 function $y : (-\varepsilon, \varepsilon) \to \mathbb{R}^{(m-s)}$ such that $y(0) = 0, \hat{C}(y(t), t) = 0$ and

$$y'(t) = -\nabla_y \hat{C}(y(t), t)^{-1} \nabla_t \hat{C}(y(t), t)$$

for all $t \in (-\varepsilon, \varepsilon)$. Hence we have

$$y'(0) = -\nabla_y \hat{C}_e(0,0)^{-1} \nabla_t \hat{C}(0,0) = -\nabla_y \hat{C}(0,0)^{-1} \nabla C_e(\overline{x}) d = 0$$

The equivalence of the blue terms is clear recalling that $\nabla_t \hat{C}_i(0) = c_i(\overline{x})d$. Now put $x(t) = \overline{x} + td + \nabla C_e(\overline{x})^T y(t)$ for all $t \in (-\varepsilon, \varepsilon)$. It is clear that $x(t)$ satisfies all desired properties. In particular $x \in C^1, x(0) = \overline{x}, x'(0) = d$ and $c_i(x(t)) = 0$ for all $t \in (-\varepsilon, \varepsilon)$. To see the last point recall our C^1 function satisfies $\hat{C}(y(t), y)$ which necessarily implies all its components $\hat{C}_i(y(t), y) = c_i(x(t)) = 0$. □

Fact 53

Upon reflection you will see why we defined $\hat{C}_i(y, t)$ as above. Consider

$$\hat{C}_i(t) = c_i(\overline{x} + td) \quad i = s+1, \dots, m$$

the objective is to have for any feasible direction d a perturbation td such that $\hat{C}_i(t) = 0$(that is still satisfies constraint, so perturbation is contained within feasible set).Now an intuitive thing to do is to introduce a "correction term"

$$\hat{C}_i(y, t) = c_i(\overline{x} + td + \nabla C_e(\overline{x})^T y) \quad i = s+1, \dots, m$$

because we know with implicit function we can find a local area in which

$$\hat{C}(y(t), t) = 0$$

as desired

Theorem 54

Let $\overline{x}$ be feasible for NLP. Then the following implications hold

$$\text{LICQ}(\overline{x}) \to \text{MFCQ} \to \Omega \text{ regular}$$

Proof. Let $A(\overline{x})$ represent

$$\begin{pmatrix} \nabla c_i(\overline{x})^T & (i \in I(\overline{x}) \\ \nabla c_i(\overline{x})^T & (i = s+1, \dots, m) \end{pmatrix} \in \mathbb{R}^{|I(\overline{x}) + (m-s)| \times n}$$

Then define

$$A(\overline{x})d = \begin{pmatrix} -e \\ 0 \end{pmatrix}$$

with $e \in \mathbb{R}^{|I(\overline{x})|}$ being the vector of all ones. Since we know that $A(\overline{x})$ is non singular, there exists a solution d which clearly satisfies the MFCQ(just take the matrix product and see). To show regularity it suffices to show that

$$L(\overline{x}) = \left\{ d : \nabla c_i(\overline{x})^T d \leq 0, i \in I(\overline{x})^T d = 0, i = s+1, \ldots, m \right\} \subset T(\overline{x}|\Omega)$$

However we know that by 49 there exist a C^1curve $x : (-\varepsilon, \varepsilon) \to \mathbb{R}^n$ such that $x(t) \in \Omega$ and $x'(0) = d, x(0) = \overline{x}$. Then but we also know $x'(0) = d \in T(\overline{x|\Omega})$ by definition since

$$d(t) = x'(0) = \lim_{k \to} \frac{x(t_k) - \overline{x}}{t_k - 0}$$

where the points $x(t_k) \in \Omega$.

6 conic optimization(6)

6.1 optimality conditions and conic duality

So far we have learnt two optimality constraints that prove duality namely

1. normal cones for linear programming strong duality
2. slater condition for convex optimization strong duality(see `Stanford Convex Analysis notes`)

We will now add to our repertoire of strong duality conditions: by learning about *conic* duality...to be continued

7 Gradient descent(7)

We now start exploring first-order optimization methods for nonlinear optimization. We begin our exploration in the case of minimization of *unconstrained* differentiable functions that is optimization problems of the form

$$\min_x f(x)$$

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

7.1 the fundamental idea of gradient descent

Fact 55

Suppose that $f(x)$ can be approximated by 1st order methods that is using taylor expansion we have

$$f(x) \approx f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$$

for all x "close to" x_t The idea of gradient descent is to move in the direction that minimizes the approximation of the objective above, that is, move a certain amount $\eta > 0$ in the direction $-\nabla f(x_t)$ of steepest descent of the function. That is we recursively define x_t to be

$$x_{t+1} = x_t - \eta \nabla f(x_t)$$

We call x_0 the initial point and the parameter $\eta > 0$ the "stepsize" or "learning rate"

7.2 Analysis for L-smooth functions

Specifically we will require that the gradient $\nabla f(x)$ be L-Lipschiz continuous for some constant $L \geq 0$. This condition is often called **L-smoothness** in the literature. We will consider the case for general functions with arbitrary convex domains Ω but this section we only care for the case $\Omega = \mathbb{R}^n$

Definition 56 (L-smoothness)

A differentiable function $f : \Omega \to \mathbb{R}$ is **L-smooth** if its gradient is L-Lipschitz continuous that is

$$||\nabla f(x) - \nabla f(y)||_2 \leq L\,||x - y||_2 \quad \forall x, y \in \Omega$$

An immediate consequence of this is that the function admits a **quadratic upper bound**.

Theorem 57 (Quadratic upper bound)

Let $f : \Omega \to \mathbb{R}$ be L-smooth on a convex domain Ω. Then we can upper bound the function f as

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\,||y - x||_2^2 \quad \forall x, y \in \Omega$$

Proof. Express the growth $f(y) - f(x)$ as the integral of the gradient on the line connecting x to y ten use the lipschitzness bound on the growth of the gradient ∇f to bound the growth

$$\begin{aligned}
f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t \cdot (y - x)), y - x \rangle dt \\
&= \left(\int_0^1 \langle \nabla f(x + t \cdot (y - x)) - \nabla f(x), y - x \rangle dt \right) + \langle \nabla f(x), y - x \rangle \\
&\leq \left(\int_0^1 ||\nabla f(x + t \cdot (y - x)) - \nabla f(x)||_2 \cdot ||y - x||_2 \, dt \right) + \langle \nabla f(x), y - x \rangle \\
&\leq \left(\int_0^1 tL\,||y - x||_2^2 \, dt \right) + \langle \nabla f(x), y - x \rangle \\
&= \frac{L}{2}\,||y - x||_2^2 + \langle \nabla f(x), y - x \rangle
\end{aligned}$$

where the 3rd line follows by **Cauchy inequality** for pre-hilbert spaces(aka inner product spaces that are not necessarily complete) and the 2nd last line follows by direction application of 56

Theorem 58

For twice differentiable functions $f : \Omega \to \mathbb{R}$ defined on an open set $\Omega \subseteq \mathbb{R}^n$ an equivalent condition for L-smoothness is

$$\left|v^T \nabla^2 f(x) v\right| \leq L \quad \forall x \in \Omega, v \in \mathbb{R}^n : ||v||_2 = 1$$

Proof. This should look very familiar! Recall 24. Now pick any two $x, y \in \Omega$ and $t \in (0, 1]$. Define $x_t = x + t \cdot (y - x)$ where $x_t \in \Omega$ by definition of convexity of Ω just like before. From the quadratic upper bound we can write

$$f(x_t) \leq f(x) + \langle \nabla f(x), x_t - x \rangle + \frac{L}{2} ||x_t - x||_2^2$$
$$f(x) \leq f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{L}{2} ||x_t - x||_2^2$$

By bilearity of the inner product we can also write

$$f(x) \leq f(x_t) + \langle -\nabla f(x_t), x_t - x \rangle + \frac{L}{2} ||x_t - x||_2^2$$

then summing the inequalities we have

$$\begin{aligned}
0 &\leq \langle \nabla f(x) - \nabla f(x_t), x_t - x \rangle + L\, ||x_t - x||_2^2 \\
&= \langle \nabla f(x) - \nabla f(x_t), t \cdot (y - x) \rangle + L\, ||t(y - x)||_2^2 \\
&= t \cdot \langle \nabla f(x) - \nabla f(x_t), y - x \rangle + t^2 L\, ||x_t - x||_2^2
\end{aligned}$$

where we brought t out by homogeneity of the norm Dividing both sides by t^2 we have

$$\frac{\langle \nabla f(x + t \cdot (y - x)) - \nabla f(x), y - x \rangle}{t} + L\, ||y - x||_2^2$$

then multiplying the 1st argument of the inner product and term with L by -1(essentially negating both sides and changing the sign of the inquality accordingly) and taking the limit as $t \downarrow 0$ we therefore have

$$\langle (y - x), \nabla^2 f(x)(y - x) \rangle \leq L\, ||y - x||_2^2$$

Because Ω is open, that means that every x has a neighborhood contained in Ω(i.e x is an interior point). And because the choice of y arbitrary we are able to choose points y all around the point x and therefore the direction $y - x$ is abirtrary as well. This implies the hessian matrix $\nabla^2 f(x)$ is **negative semi-definite** with respect to $L\, ||y - x||_2^2$

$$\nabla^2 f(x) \preccurlyeq L\, ||y - x||_2^2$$

Now then upon normalizing $y - x$ to get v the theorem clearly follows

Theorem 59 (Gradient descent lemma)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth. Then for any $0 < \eta \leq \frac{1}{L}$ each step of gradient descent as described in 55 gaurantees

$$f(x_{t+1}) \leq f(x_t) - \frac{\eta}{2} ||\nabla f(x_t)||_2^2$$

Proof. First using **quadratic upperbound** 57 for the choice $x = x_t, y = x_{t+1}$

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} ||x_{t+1} - x_t||_2^2$$

Then plugging in the gradient descent step $x_{t+1} = x_t - \eta\nabla f(x_t)$ we therefore obtain

$$f(x_{t+1}) \leq f(x_t) - \eta\,||\nabla f(x_t)||_2^2 + \frac{L}{2}\eta^2\,||\nabla f(x_t)||_2^2 = f(x_t) - \eta\left(1 - \frac{L}{2}\eta\right)||\nabla f(x_t)||_2^2$$

If $\eta \leq \frac{1}{L}$ as assumed in the theorem then $\left(1 - \frac{L}{2}\eta\right) \geq \frac{1}{2}$ in which case the result follows □

The previous result shows that for L-smooth functions there exists a good choice of learning rate (namely $\eta = \frac{1}{L}$) such that each step of gradient descent guarantees to improve the function value if the current point does not have a zero gradient.

Note that the gradient descent lemme then also implies that if f is lower bounded then the *gradients* of the points produced by gradient descent must eventually become small.

Theorem 60

Suppose f is bounded below. Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smoth an let $0 \leq \eta \leq \frac{1}{L}$. Then by running gradient descent for T interations at least one of the points $\{x_t\}$ encountered must satisfy

$$||\nabla f(x_t)||_2 \leq \sqrt{2 \cdot \frac{f(x_0) - f_\star}{\eta T}}$$

Remark 61. *Notice that such a condition did not require the convexity of f*

Proof. Let $f_\star = \inf f$ which exists since lower bounded. Consider running the algorithm for T iterations. Then by induction on the **gradient descent lemma** 59 we have

$$f_\star \leq f(x_T) \leq f(x_0) - \frac{\eta}{2}\sum_{t=0}^{T-1}||\nabla f(x_t)||_2^2$$

Hence we have

$$\sum_{t=0}^{T-1}||\nabla f(x_t)||_2^2 \leq \frac{2}{\eta}(f(x_0) - f_\star)$$

Then dividing all by $T - 1$ we have

$$\min_t ||\nabla f(t)||_2^2 \leq \frac{1}{T}\left(\sum_{t=0}^{T-1}||\nabla f(x_t)||_2^2\right) \leq \frac{2}{\eta T}(f(x_0) - f_\star)$$

the 1st equality follow by definition of an arithmetic mean. Thefore we have

$$\min_t ||\nabla f(t)||_2^2 \leq \frac{2}{\eta T}(f(x_0) - f_\star)$$

Remark 62. *Think about how this implies the theorem. If we had* $\max_t ||\nabla f(t)||_2^2 \leq \frac{2}{\eta T}(f(x_0) - f_\star)$ *then all t will satisfy this bound. But we have* min *so in the worst case at least one t which corresponds to the minimum value of* $||\nabla f(x)||$ *satisfies the bound*

And therefore taking square roots, the theorem follows □

To recap so far we have proven that for an L smooth function with a good choice of learning rate, we are guaranteed that each step of gradient descent does get a more optimal value and if f is lower bounded, the gradient produced eventually becomes small. But we have yet to address if that means the value of f does become small as well as desired. For example a small gradient indicates a condition of local optimality as it means you are almost a critical point. But that does not imply global optimality.

Example 63

Consider the increasing function

$$f(x) = \varepsilon \log(1 + e^x)$$

Then

$$\nabla f(x) = \varepsilon \underbrace{\frac{1}{\ln 10}\left(\frac{e^x}{1+e^x}\right)}_{<1} < \varepsilon \quad \forall x \in \mathbb{R}$$

But clearly x may be arbitrarily far from the optimal(smallest value) while the gradient is small.

To give gaurantees in function value we will now futher assume that f is *convex*

Theorem 64 (Euclidean mirror descent lemma)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable function. Then for any choice of stepsize η any two consecutive points (x_t, x_{t+1}) produced by the gradient descent algorithm satisfy

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\left(||y - x_t||_2^2 - ||y - x_{t+1}||_2^2 + ||x_{t+1} - x_t||_2^2\right) \quad \forall y \in \mathbb{R}^n$$

Proof. The proof rests on the following critical observation, often called the *three-point equality*, which can be checked by expanding the squared norms: See that

$$\langle x_t - x_{t+1}, y - x_t \rangle = \langle x_t, y \rangle - \|x_t\|^2 - \langle x_{t+1}, y \rangle + \langle x_{t+1}, x_t \rangle$$

Next, we want to match this with the expression in terms of squared norms. Let's expand the squared norms that we'll substitute into this expression.

$$\|y - x_t\|^2 = \langle y - x_t, y - x_t \rangle = \|y\|^2 - 2\langle y, x_t \rangle + \|x_t\|^2$$

$$-\|y - x_{t+1}\|^2 = -\langle y - x_{t+1}, y - x_{t+1} \rangle = -\|y\|^2 + 2\langle y, x_{t+1} \rangle - \|x_{t+1}\|^2$$

$$\|x_{t+1} - x_t\|^2 = \langle x_{t+1} - x_t, x_{t+1} - x_t \rangle = \|x_{t+1}\|^2 - 2\langle x_{t+1}, x_t \rangle + \|x_t\|^2$$

Then

$$\langle x_t - x_{t+1}, y - x_t \rangle = -\frac{1}{2}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2\right).$$

Since f is convex and differentiable by 24 we have

$$f(y) \geq f(x_t) + \langle \nabla f(x_t), y - x_t \rangle.$$

Plugging in $x_{t+1} = x_t - \eta \nabla f(x_t)$, we then obtain

$$f(y) \geq f(x_t) + \frac{1}{\eta}\langle x_t - x_{t+1}, y - x_t \rangle.$$

Substituting the critical observation above and rearranging, we obtain

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2 + \|x_{t+1} - x_t\|_2^2\right) \quad \forall y \in \mathbb{R}^n,$$

which is the statement. □

We can now use the euclidean mirror descent lemma above to get the rate of decrease in terms of the objective function value $f(x_t)$.

First, since $x_{t+1} - x_t = -\eta \nabla f(x_t)$, we can rewrite the above as

$$f(x_t) \leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|_2^2 \quad \forall y \in \mathbb{R}^n.$$

Assuming f is L-smooth and $\eta \leq \frac{1}{L}$, we can then use the gradient descent lemma 59 to bound

$$\frac{\eta}{2}\|\nabla f(x_t)\|_2^2 \leq f(x_t) - f(x_{t+1}),$$

obtaining the following corollary via

$$\begin{aligned} f(x_t) &\leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2\right) + \frac{\eta}{2}\|\nabla f(x_t)\|_2^2 \\ &\leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2\right) + f(x_t) - f(x_{t+1}) \end{aligned}$$

Corollary 65

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth, and $0 < \eta \leq \frac{1}{L}$. Then, any two consecutive points (x_t, x_{t+1}) produced by gradient descent satisfy

$$f(x_{t+1}) \leq f(y) + \frac{1}{2\eta}\left(\|y - x_t\|_2^2 - \|y - x_{t+1}\|_2^2\right) \quad \forall y \in \mathbb{R}^n.$$

□

In particular, if f has a minimizer x_*, the above corollary implies that, whenever $\eta \leq \frac{1}{L}$,

$$f(x_{t+1}) \leq f(x_*) + \frac{1}{2\eta}\left(\|x_* - x_t\|_2^2 - \|x_* - x_{t+1}\|_2^2\right).$$

Summing over both sides by $t = 0, \ldots, T-1$, and noticing that the right-hand side is telescopic(middle terms in sum cancel), we have

$$\sum_{t=0}^{T-1} f(x_{t+1}) \leq Tf(x_*) + \frac{1}{2\eta}\|x_* - x_0\|_2^2 - \frac{1}{2\eta}\|x_* - x_T\|_2^2 \leq Tf(x_*) + \frac{1}{2\eta}\|x_* - x_0\|_2^2$$

Since $f(x_t)$ is nonincreasing in t by the gradient descent lemma(every iteration results in an improvement of the value of $f(x_t)$), then $\sum_{t=0}^{T-1} f(x_{t+1}) \geq Tf(x_T)$, and so we can write

$$Tf(x_T) \leq Tf(x_*) + \frac{1}{2\eta}\|x_* - x_0\|_2^2 \quad \Longrightarrow \quad f(x_T) \leq f(x_*) + \frac{1}{2T\eta}\|x_* - x_0\|_2^2.$$

So, we have proved the following.

Theorem 66

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and L-smooth with minimizer $x_* \in \mathbb{R}^n$ and $0 < \eta < \frac{1}{L}$. The t-th point $x_t \in \mathbb{R}^n$ produced by gradient descent satisfies

$$f(x_t) - f(x_*) \leq \frac{\|x_* - x_0\|_2^2}{2t\eta} \quad \forall t = 1, 2, 3, \ldots$$

which implies a monotonic decrease in euclidean distance to optimality. Intuitively you know this can only be true for convex functions(just imagine if there is a hill near the optimal value) which we have proved rigorously earlier. □

...to be continued

8 Acceleration and momentum(8)

8.1 A tale of two descent modes: a second look at the descent lemmas

To recap from the previous lecture we have that for a convex and L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ where $x_\star \in \mathbb{R}^n$ is a minimizer of f and x_t ar iterates produced by gradient descent with step size $\eta > 0$,

- the **gradient descent lemma** asserts that the progress made in the *function value* in two consecutive interates x_t and x_{t+1} is at least as big as the *norm of the gradient* of f at x_t provided $\eta \leq \frac{1}{L}$ then

$$(f(x_{t+1}) - f_\star) \leq (f(x_t) - f_\star) - \frac{\eta}{2} ||\nabla f(x_t)||_2^2$$

- the **euclidean mirror descent lemma** asserts that

$$||x_{t+1} - x_\star||_2^2 \leq ||x_t - x_\star||_2^2 + \eta^2 ||\nabla f(x_t)||_2^2 - 2\eta \cdot (f(x_t) - f_\star)$$

  noting that $x_{t+1} = x_t - \eta \nabla f(x_t)$(by definition of the gradient descent step). hence the euclidean mirror descent lemma establishes that the *distance from the optimal solution* decreases fast when the optimality gap $f(x_t) - f_\star$ is large and the gradient norm $||\nabla f(x_t)||_2$ is small

and on combination of these 2 lemmas we arrive at 66 which implies optimality in function value is reached as the euclidean distance to optimality decreases. To summarize further we can say that these 2 lemmas focus on the following performance metrics

- the gradient descent lemma focuses on the progress in the function value
- while the euclidean mirror descent lemma focuses on progress on the euclidean distance to optimality

We now want to explore another useful performance metric - the speed at which optimality is found(i.e the number of iterates required to arrive at optimality). In particular we want to construct a form of accelerated descent. To that end we first consider a thought experiment: imagine running gradient descent on an L-smooth and convex function using stepsize η. We now consider two extreme cases

- **large gradients**

Theorem 67

Suppose that all gradients of the points produced by gradient descent satisfy

$$||\nabla f(x_t)||_2^2 \geq \gamma \quad \text{at all } t = 0, 1 \dots$$

for some constant $\gamma > 0$. In this case using the stepsize $\eta = \frac{1}{L}$, after $T_{\text{half}} := \frac{L}{\gamma}(f(x_0) - f_\star)$ iterations the optimality gap will halve that is

$$f(x_{T_{\text{half}}}) - f_\star \leq \frac{f(x_0) - f_\star}{2}$$

Proof. Since our function is L-smooth, the gradient descent lemma implies that

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2L} ||\nabla f(x_t)||_2^2 \leq -\frac{\gamma}{2L}$$

so after T_{half} iterations that is

$$\sum_{t=0}^{T_{\text{half}}} f(x_{t+1}) - f(x_t) \leq -\frac{\gamma}{2L} T_{\text{half}}$$

where we obtain upon subbing in our definition of T_{half} in the statement

$$f(x_{T_{\text{half}}}) - f(x_0) \leq -\frac{\gamma}{2L} T_{\text{half}} = -\frac{1}{2}(f(x_0) - f_\star)$$

where upon rearrangment yields the statement as desired

small gradients

Theorem 68

Consider the case where

$$||\nabla f(x_t)||_2^2 \leq \gamma \quad \text{at all } t = 0, 1, \dots$$

for some constant $\gamma > 0$. In this case, using the stepsize $\eta := (f(x_0) - f_\star)/2\gamma$ after

$$T_{\text{half}} := 4\gamma \frac{||x_0 - x_\star||_2^2}{(f(x_0) - f_\star)^2}$$

interations we will find a point $x_{\text{half}} \in \{x_0, x_1, \dots, x_{T_{\text{half}}}\}$ with optimality gap

$$f(x_{\text{half}}) - f_\star \leq \frac{f(x_0) - f\star}{2}$$

Proof. In this case for every $\eta > 0$, since our function is differentiable(since L-smooth) and convex the euclidean mirror descent guarantees that

$$||x_{t+1} - x_\star||_2^2 \leq ||x_t - x_\star||_2^2 + \eta^2 ||\nabla f(x_t)||_2^2 - 2\eta \cdot (f(x_t) - f_\star)$$

where $x_{t+1} - x_t = -\nabla f(x_t)$ so upon rearrangement with have

$$f(x_t) - f_\star \leq \frac{1}{2\eta} \left(||x_t - x_\star||_2^2 + \eta^2 ||\nabla f(x_t)||_2^2 - ||x_{t+1} - x_\star||_2^2 \right)$$
$$f(x_t) \leq f_\star + \frac{1}{2\eta} ||x_t - x_\star||_2^2 + \frac{\eta}{2} ||\nabla f(x_t)||_2^2 - \frac{1}{2\eta} ||x_{t+1} - x_\star||_2^2$$

now taking the sum over T interations starting from $t = 0$ to $t = T - 1$ then dividing by T we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} f(x_t) &\leq f_\star + \frac{1}{2\eta T}\,||x_0 - x_\star||_2^2 + \frac{\eta}{2T}\sum_{t=0}^{T-1}||\nabla f(x_t)||_2^2 - \frac{1}{2\eta T}\,||x_T - x_\star||_2^2 \\ &\leq f_\star + \frac{1}{2\eta T}\,||x_0 - x_\star||_2^2 + \frac{\eta}{2T}\sum_{t=0}^{T-1}||\nabla f(x_t)||_2^2 \\ &< f_\star + \frac{1}{2\eta T}\,||x_\star - x_0||_2^2 + \frac{n\gamma}{2}\end{aligned}$$

now subbing value of η defined above and letting $T = T_{\text{half}}$ which is also as defined above,we have

$$\begin{aligned}\frac{1}{T_{\text{half}}}\sum_{t=0}^{T_{\text{half}}-1}(f(x_t) - f_\star) &< \frac{\gamma}{T_{\text{half}}(f(x_0) - f_\star)}\,||x_\star - x_0||_2^2 + \frac{n\gamma}{2} \\ &= \frac{f(x_0) - f_\star}{2}\end{aligned}$$

However knowing that the minimum is bounded by the average we have that

$$\min_{t=0}^{T_{\text{half}}-1} f(x_t) - f_\star < \frac{f(x_0) - f_\star}{2}$$

but we can increase to T_{half} a minimum can only get smaller when the set is expanded which will still satisfy the upper bound. That is

$$\min_{t=0}^{T_{\text{half}}} f(x_t) - f_\star < \frac{f(x_0) - f_\star}{2}$$

Balancing the two cases so suppose the stepsize is chosen well(i.e $0 < \eta < \frac{1}{L}$). In the first case $T_{\text{half}} \propto \frac{1}{\gamma}$ while in the second $\propto \gamma$. That is as γ increases the required iterations in the first case decreases strictly while in the second it increases strictly.To see why strict, see their definitions, the other terms that multiply γ in their respective defintions are all positive. Therefore the value of γ that minimizes the maximum halving time across both cases can only be when they are equal(since its the only intersection point anyway if you think about it graphically) so solving

$$\frac{L}{\gamma}(f(x_0) - f_\star) = 4\gamma\frac{||x_0 - x_\star||_2^2}{(f(x_0) - f_\star)^2} \quad \rightarrow \quad \gamma = \frac{\sqrt{L}(f(x_0) - f_\star)^{3/2}}{2\,||x_0 - x_\star||}$$

then on subtitution into their respective equations for T_{half} we find that both give

$$T_{\text{half}} = \frac{2\,||x_0 - x_\star||_2\,\sqrt{L}}{\sqrt{f(x_0) - f_\star}}$$

now compare this to when we didn't consider setting any bounds on $||\nabla f(x_t)||_2^2$ and just use the euclidean mirror descent and gradient descent lemma, in particular 66 which implies

$$f(x_t) - f_\star \leq L\frac{||x_0 - x_\star||_2^2}{2t}$$

when say $\eta = \frac{1}{L}$ which on solving when we have $f(x_t) - f_\star \leq \frac{f(x_0) - f_\star}{2}$ we get

$$t = L\frac{||x_0 - x_\star||_2^2}{f(x_0) - f_\star}$$

which is a quadratic blowup in compared to T_{half} we found earlier.

8.2 Allen-Zhu and Orrecchia Linear coupling

We now formalize our findings above since it might be the case that neither $||\nabla f(x_t)||_2^2 \geq \gamma$ at all times nor $||\nabla f(x_t)||_2^2 \leq \gamma$ at all times i.e the bound values are nt necessarily fixed with each iteration. Similarly considering "large" and "small" gradients

Proposition 69 (Allen-Zhu and Orrechia Linear Coupling)
Allen-Zhu and Orrechia proposed the following algorithm which keeps tracks of 3 sequences $\{x_t\}, \{y_t\}, \{z_t\}$. The sequence x_t corresponds to the "final" iterate while the sequences y_t and z_t correspond to the short("large gradients) and long steps("small gradients") respectively.At the beginning,

$$x_0 = y_0 = z_0$$

at each iteration we let

$$\begin{aligned} x_{t+1} &:= (1-\tau)y_t + \tau z_t \quad \text{(interpolation with coupling rate } \tau) \\ y_{t+1} &:= x_{t+1} - \frac{1}{L}\nabla f(x_{t+1}) \quad \text{("short" gradient step} \\ z_{t+1} &:= z_t - \alpha \nabla f(x_{t+1}) \quad \text{("long" gradient step, with stepsize } \alpha) \end{aligned}$$

Definition 70
The quantity $z_t = z_0 - \alpha \sum_{s=1}^{t} \nabla f(x_s)$ keeps tracks of past gradients which is then combined into the definition of x_{t+1}. This term is often called **momentum**

Theorem 71 (coupled euclidean mirror descent lemma)
At all times t

$$f(x_{t+1}) - f_\star \leq \frac{1}{2\alpha}\left(||x_\star - z_t||_2^2 - ||x_\star - z_{t+1}||_2^2 + ||z_t - z_{t+1}||_2^2\right) + \frac{1-\tau}{\tau}(f(y_t) - f(x_{t+1})$$

Proof. Using the **three point equality** as covered in previous sections we write

$$\langle z_t - z_{t+1}, z_t - x_\star \rangle = \frac{1}{2}\left(||x_\star - z_t||_2^2 - ||x_\star - z_{t+1}||_2^2 + ||z_t - z_{t+1}||_2^2\right)$$

using the fact that $z_{t+1} = z_t - \alpha \nabla f(x_{t+1})$ we can therefore write

$$\alpha\langle \nabla f(x_{t+1}), z_t - x_\star \rangle = \frac{1}{2}\left(||x_\star - z_t||_2^2 - ||x_\star - z_{t+1}||_2^2 + ||z_t - z_{t+1}||_2^2\right)$$

now using the linear coupling proposition above we have that

$$x_{t+1} = (1-\tau)y_t + \tau z_t \quad \Rightarrow \quad (x_{t+1} - x_\star) = (z_t - x_\star) + \frac{1-\tau}{\tau}(y_t - x_{t+1})$$

so with this consider that ...to be continued

Theorem 72

Let α and τ be defined so that

$$\alpha := \frac{||x_\star - z_0||_2}{\sqrt{2(f(x_0) - f_\star)}} \quad \tau := \frac{1}{1 + \alpha L}$$

Then the **Allen Zhu and Orrechia accelerated gradient descent** finds at least one iterate x_t such that

$$f(x_t) - f_\star \leq \frac{1}{2}(f(x_0) - f_\star)$$

within

$$T_{\text{half}} := \frac{2\,||x_\star - x_0||_2\,\sqrt{2L}}{\sqrt{f(x_0) - f_\star}}$$

...to be continued

9 Projected gradient descent and mirror descent(9)

10 Stochastic gradient descent(10)

10.1 stochastic gradient descent lemma

Fact 73 (stochastic gradient descent)

Evaluating the exact gradient ∇f can be computationally expensive. So we use a cheap unbiases estimator $\tilde{\nabla} f$ of it where

$$\mathbb{E}_t[\tilde{\nabla} f(x_t)] = \nabla f(x_t)$$

Therefore we now have the SGD algorithm:

$$\boxed{x_{t+1} := x_t - \eta \tilde{\nabla} f(x_t)}$$

Recall that an L-smooth function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the quadratic upper bound

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\,||x - y||_2^2\,, \forall x, y \in \mathbb{R}^n$$

Then knowing that $x_{t+1} = x_t - \tilde{\nabla} f(x_t)$ by defintion the SGD step we have that letting $y = x_{t+1}$ and $x = x_t$ above

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\,||x_t - x_{t+1}||_2^2 = f(x_t) - \eta \langle \nabla f(x_t), \tilde{\nabla} f(x_t) \rangle + \frac{L}{2}\eta^2\,||\tilde{\nabla} f(x_t)||_2^2$$

now taking conditional expectations on both sides we have proven

Theorem 74 (Stochastic Gradient descent lemma)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-smooth and $\eta > 0$ be an arbitrary stepsize. Two consecutive iterates (x_t, x_{t+1}) produced by the stochastic gradient descent algorithm above satisfy

$$\mathbb{E}_t[f(x_{t+1})] \leq f(x_t) - \eta\,||\nabla f(x_t)||_2^2 + \frac{L}{2}\eta^2 \mathbb{E}_t[||\tilde{\nabla} f(x_t)||_2^2]$$

the quantity $\mathbb{E}_t[\left|\left|\tilde{\nabla} f(x_t)\right|\right|_2^2]$ controls the quadratic term in the previous bound and is often called the **variance** of the stochastic gradient $\tilde{\nabla} f$

Just like the gradient descent lemma for exact gradient, the *stochastic gradient descent lemma* guarantees descent in function value in expectation, when $\eta > 0$ is sufficiently small. Specifically suppose that $\mathbb{E}_t[\left|\left|\nabla \tilde{f(x_t)}\right|\right|_2^2] \leq G$ at all times t(we won't be having a discussion if the quadratic term wasn' bounded). Then the gradient descent lemma can be written as

$$||f(x_t)||_2^2 \leq \frac{1}{\eta}\mathbb{E}_t[f(x_t) - f(x_{t+1})] + \frac{L}{2}\eta G$$

then summing over all times $t = 0, 1 \ldots, T-1$ we therefore obtain

$$\sum_{t=0}^{T-1} ||\nabla f(x_t)||_2^2 \leq \frac{1}{\eta}\left(\sum_{t=0}^{T-1} \mathbb{E}_t[f(x_t) - f(x_{t+1})]\right) + \frac{L}{2}\eta G T$$

taking expectations of both sides we see that...to be continued

11 Hessians,preconditioning and Newton's Method(12)

So far we have been concerned with first order methods, that is, those optimization methods that use gradient information. Today we start discussing *second order methods*. For now we restrict our attention to problems of the form

$$\min_x f(x)$$

where $x \in \mathbb{R}^n$ and $f(x)$ is *twice-differentiable* function

11.1 from first order to second order taylor approximations

Fact 75

Recall that the second order Taylor expansion of a function $f(x)$ around a point x_t is given by

$$f(x) \approx f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2}\langle x - x_t, \nabla^2 f(x_t), (x - x_t) \rangle$$

where $\nabla^2 f(x_t)$ is the Hessian matrix of f at x_t

The minimum of $f(x)$ can be found in closed form by setting the gradient (with respect to x) of the above expression to zero that is

$$\nabla_x \left(f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2}\langle x - x_t, \nabla^2 f(x_t), (x - x_t) \rangle \right) = 0$$

which gives

$$\nabla f(x_t) + \nabla^2 f(x_t)(x - x_t) = 0 \quad \Rightarrow \quad x = x_t - [\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$$

so we see that by moving from first order to second order Taylor approximation and assuming that $\nabla^2 f(x_t)$ is invertible we the natural dirction of descent changes from

$$\underbrace{d = -\nabla f(x_t)}_{\text{first order}} \quad \text{to} \quad \underbrace{d = -[\nabla^2 f(x_t)]^{-1} \nabla f(x_t)}_{\text{second order}}$$

11.2 from gradient descent to newton's method

having established the second order direction of descent we can define the natural generalization of nth gradient descent algorithm to the second order setting.

Definition 76

The algorithm which takes the same *damped Newton's method* is given by the update rule

$$\boxed{x_{t+1} = x_t - \eta[\nabla^2 f(x_t)]^{-1}\nabla f(x_t)}$$

is known as **Newton's method**

Example 77

Consider

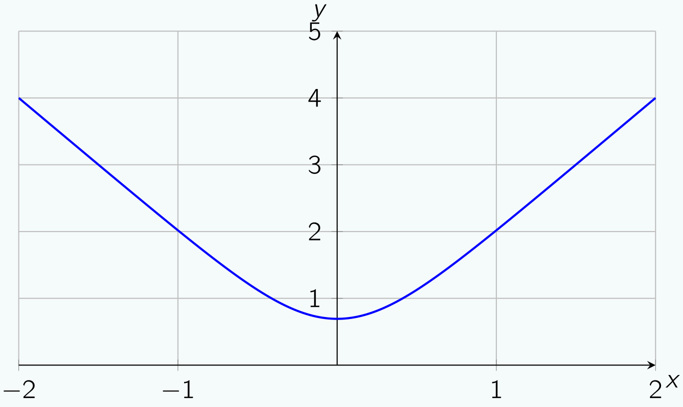
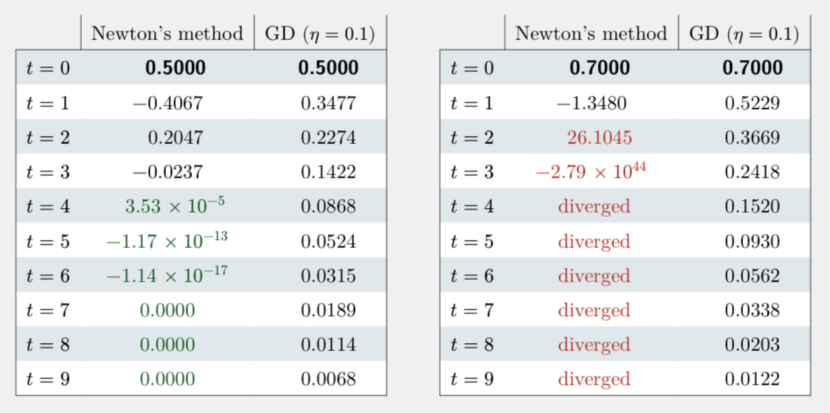


Figure 7: $f(x) = \log(e^{2x} + e^{-2x})$

Newton's method converged very quicly when we started at $x_0 = 0.5$ but diverges when starting at $x_0 = 0.7$ as reflected in



	Newton's method	GD ($\eta = 0.1$)
$t = 0$	**0.5000**	**0.5000**
$t = 1$	-0.4067	0.3477
$t = 2$	0.2047	0.2274
$t = 3$	-0.0237	0.1422
$t = 4$	3.53×10^{-5}	0.0868
$t = 5$	-1.17×10^{-13}	0.0524
$t = 6$	-1.14×10^{-17}	0.0315
$t = 7$	0.0000	0.0189
$t = 8$	0.0000	0.0114
$t = 9$	0.0000	0.0068

	Newton's method	GD ($\eta = 0.1$)
$t = 0$	**0.7000**	**0.7000**
$t = 1$	-1.3480	0.5229
$t = 2$	26.1045	0.3669
$t = 3$	-2.79×10^{44}	0.2418
$t = 4$	diverged	0.1520
$t = 5$	diverged	0.0930
$t = 6$	diverged	0.0562
$t = 7$	diverged	0.0338
$t = 8$	diverged	0.0203
$t = 9$	diverged	0.0122

How? Why? these questions will be answered in the following subsections

11.3 Analysis of Newton's method

Theorem 78

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with invertible Hessian and let $x_\star$ be a local minimum of f. The distance to optimality of the iterates x_t generated by the damped Newton's method with stepsize $\eta > 0$ satisfy

$$x_{t+1} - x_\star = (I - \eta H_t)(x_t - x_\star)$$

where

$$H_t := [\nabla^2 f(x_t)]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x_t - x_\star)) d\lambda$$

Proof. Consider

$$\begin{aligned}
[\nabla^2 f(x_t)]^{-1} \nabla f(x_t) &= [\nabla^2 f(x_t)](\nabla f(x_t) - \nabla f(x_\star)) \quad \text{since } \nabla f(x_\star) = 0 \\
&= [\nabla^2 f(x_t)]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x_t - x_\star))(x_t - x_\star) d\lambda \\
&= \left([\nabla^2 f(x_t)]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x_t - x_\star)) d\lambda \right) (x_t - x_\star) \\
&= H_t (x_t - x_\star)
\end{aligned}$$

then on substitution into the update rule of damped newton method we find

$$x_{t+1} - x_\star = (x_t - x_\star) - \eta [\nabla^2 f(x_t)]^{-1} \nabla f(x_t) = (I - \eta H_t)(x_t - x_\star)$$

as desired □

As a first corollary of the previous lemma we can derive a convergence gaurantee for Newton's method when starting from a point that is "close enough" to a minimum with sufficient curvature. This is what is known as a **local covergence gauarantee**. To that end we make the 2 following assumptions

1. In particular let $x_\star$ be a local minimum of f with strong curvature that is a point such that

   $$\nabla f(x_\star) = 0, \quad \text{and} \quad \nabla^2 f(x_\star) \succcurlyeq \mu I$$

   for some $\mu > 0$.

2. Furthermore assume that f is *smooth* in the sesne that its Hessian is **M-Lipschitz continuous** that is

   $$\left|\left| \nabla^2 f(x) - \nabla^2 f(y) \right|\right|_s \leq M \cdot ||x - y||_2$$

Theorem 79

Under the 2 assumptions above the spectral norm(see appendix) of the matrix $I - H_t$ induced at time t by the iterate x_t produced by Newton's method satisfies

$$||I - H_t||_s \leq \frac{M}{\mu} ||x_t - x_\star||_2$$

whenever

$$||x_t - x_\star|| \leq \frac{\mu}{2M}$$

Theorem 80 (Local convergence guarantee)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with M Lipschitz continuous Hessian and let $x_\star$ be a local minimum of f with strong curvaturethat is a point such that

$$\nabla f(x_\star) = 0 \quad \text{and} \quad \nabla^2 f(x_\star) \succcurlyeq \mu I$$

for some $\mu > 0$. Then as long as we start Newton's method from a point x_0 with distance

$$||x_0 - x_\star|| \leq \frac{\mu}{2M}$$

from the local minimum the distance to optimality of the iterates x_t generated by Newton's method decays as

$$\frac{||x_{t+1} - x_\star||_2}{\mu/M} \leq \left(\frac{||x_t - x_\star||_2}{\mu/M}\right)^2$$

Corollary 81 (Global convergence guarantee)
Let $f : \mathbb{R}^n$ be twice differentiable μ strong convex and L smooth. Then the distance to optimality of the iterates x_t generated by damped Newton's method with stepsize $\eta \leq \frac{\mu}{L}$ decays exponentially fast at the rate

$$||x_{t+1} - x_\star|| \leq \left(1 - \eta\frac{\mu}{L}\right) ||x_t - x_\star||_2$$

11.4 Appendix: Spectral Norm

Definition 82
The **spectral norm** of a matrix A measures the maximum increase in Euclidean norm that A can produce that is

$$||A||_s := \max_{x \neq 0} \frac{||Ax||_2}{||x||_2}$$

Theorem 83
For a symmetric matrix A the spectral norm is equal to the maximum absolute value of any eigenvalue of A

Proof. The diagonal case is straight forward since all the eigenvalues will then lie on the diagonal. To show the result in general we will show that we can reduce all cases to the diagonal case. Consider that any symmetric A can be expressed as

$$A = Q^T \Lambda Q$$

where Q is an orthogonal matrix(that is $||Qv||_2 = \left|\left|Q^T v\right|\right|_2 = ||v||_2$ for all v,aka lenght preserving under euclidean norm). Λ is a diagonal matrix with eigenvalues of A on the diagonal. We know that this decomposition exists by the **spectral theorem**(see your `functional analysis` and `Artin Algebra 1` notes). Therefore we have

$$\max_{x \neq 0} \frac{||Ax||_2}{||x||_2} = \max_{x \neq 0} \frac{\left|\left|Q^T \Lambda Q x\right|\right|_2}{||x||_2} = \max_{x \neq 0} \frac{||\Lambda Q x||_2}{||Qx||_2}$$

as desired

Corollary 84

Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then $||A||_s \leq k$ if and only if

$$-kI \succcurlyeq A \succcurlyeq kI$$

Proof. This is equivalent to asking if every eigenvalue of A is in the range $[-k, k]$ so the result follows from the previous theorem

Theorem 85

For any matrices $A, B \in \mathbb{R}^{n+m}$ we have

$$||AB||_s \leq ||A||_s \, ||B||_s$$

Proof. Let $x \neq 0$ be a vector then

$$\frac{||ABx||_2}{||Bx||_2} \leq ||A||_s = \max_{y\neq 0} \frac{||Ay||_2}{||y||_2}$$

basically Bx just corresponds to one particular value of y. Similary we have

$$||A||_s \frac{||Bx||_2}{||x||_2} \leq ||A||_s \max_{y\neq 0} \frac{||By||_2}{||y||_2} = ||A||_s \, ||B||_s$$

where x is one particular value of y here. So combining the 2 inequalities we have that

$$||ABx||_2 \leq ||A||_s \, ||Bx||_2 \leq ||A||_s \, ||B||_s \, ||x||_2$$

so we now have

$$\frac{||ABx||_2}{||x||_2} \leq ||A||_s \, ||B||_s$$

however because the value of x is arbitrary we may take the maximum over $x \neq 0$ on both sides to optain the theorem as desired

12 Adaptive preconditioning: AdaGrad and ADAM