Game Theory, Optimization, Reinforcement Learning **NTUA, Spring 2021**

Meeting 1: Convex Analysis and Optimization (A)

Advisors: ***Ioannis Panageas, Emmanouil Vlatakis*** **Student:** ***Fivos Kalogiannis***

What happens in Vegas, stays in Vegas

A popular saying on the convexity of the set of things that happen in Las Vegas

Objectives

Be able to answer the following:

1. Consider vectors $x_1, x_2, \dots \in \mathbb{R}^d$. What is a *convex combination* of such vectors?
2. What is a *convex set*?
3. What is the *epigraph* of a function?
4. Prove the equivalency of the following definitions of a convex set (hint: Jensen's inequality):
   (a) A function $f(x)$ is convex iff $\mathrm{dom} f$ is convex and $\forall x, y \in \mathrm{dom} f$ and $t \in [0, 1]$:
   $$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$
   (b) The *epigraph* of a function is a convex set.
5. Prove the following conditions for the convexity of a function:
   (a) (*First Order Condition*) A differentiable function $f(x)$ is convex iff $\mathrm{dom} f$ is a convex set *and* $\forall x, y \in \mathrm{dom} f$:
   $$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$
   (b) (*Second Order Condition*) A twice-differentiable function $f(x)$ is convex iff $\mathrm{dom} f$ is a convex set and the Hessian is positive definite:
   $$\nabla^2 f(x) \succeq 0$$
6. Given the definition of the Lipschitz Condition (write it down):
   (a) Find a function that does satisfy the Lipschitz Condition but is not continuous
   (b) Find a λ-Lipschitz function that *is* continuous
7. Define *strong convexity*
8. Write down and plot:
   - a function that is convex but not *strongly convex*
   - a strongly convex function
9. *Fenchel/convex conjugate* transform of a function
   (i) What is it? (Give a formal definition)
   (ii) Give its geometric interpretation

(iii) Why is having *every* convex conjugate of a function f is enough to define f itself?

(iv) Prove Fenchel's inequality

10. Prove that the following sentences are equivalent:

(a) f is strongly convex with parameter μ

(b) $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|y-x\|^2$

(c) $f(y) \geq f(x) + s^T(y-x) + \frac{\mu}{2}\|y-x\|^2 \ \forall x, y$ and any $s_x \in \partial f(x)$

(d) $(s_y - s_x)^T(y-x) \geq \mu\|y-x\|^2 \ \forall x, y$ and any $s_x \in \partial f(x), s_y \in \partial f(y)$

11. What is the subgradient? Define it.

1.1 Introduction

Convex analysis is the field of mathematics occupied with the concept of *convexity.* It is inherently related to the concept of mathematical optimization and its ideas and tools are omnipresent in fields like machine learning, leading to a dramatic rise of interest in the field in recent years. Convexity is –intuitively speaking– about the property of elements of a set being bound to end up in the same set after certain kinds of "interactions" amongst them. The reader should probably have by now acquainted themselves with some kind of convexity through-out their studies. Usually this incorporates thinking about convexity in terms of *convex functions* and the property of things ending up in the initial set might seem somewhat confusing. We are going to try and bridge the layman's grasp of convexity as tangent lines of a curve remaining below it at any point and the initial informal description about the property of sets whose elements will under certain "operations" result in elements belonging to the initial set.

Notation

Notation will not stray from conventions most people hold when writing mathematics. The d-dimensional Euclidean space will be signified by $\mathbb{R}^d$; scalars will be denoted with symbols like x, y, α, β, etc while random variables will favor capital letters like X and Y. Vectors (be them random or deterministic) will be written in the usual bold font: $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, etc and ℓ_p vector norms will be represented by $\|\cdot\|_p$.

1.2 Preliminaries: Convex Sets

Let's consider the set of $\mathbb{R}$ as well as the sets $\mathbb{R}^d$ where $d \in \mathbb{N}$ known. Vectors $\boldsymbol{x} \in \mathbb{R}^d$ are d-tuples $(x_1, x_2, \ldots, x_d)$. We will now move on to discuss certain concepts in the broad concept of convexity limited to such subsets of $\mathbb{R}^d$ and functions defined on them. Let's start with the elementary definition of a convex set. (We will use the words *vector* and *point* more or less interchangeably for the time being)

Definition 1 (Convex Set)**.** *A subset C of $\mathbb{R}^d$ is said to be* ***convex*** *when for every pair $\boldsymbol{x}, \boldsymbol{y} \in C \subseteq \mathbb{R}^d$ and every $\lambda \in \mathbb{R}$ for which $0 \leq \lambda \leq 1$ the following holds:*

$$\boldsymbol{z} = (1-\lambda)\boldsymbol{x} + \lambda\boldsymbol{y} \in C$$

Geometric Interpretation. A set C is called convex when for every pair of points $\boldsymbol{x}, \boldsymbol{y}$ every point $\boldsymbol{z}$ on the straight line segment defined by the pair lies within C.

Theorem 1. The intersection of convex sets is a convex set.

Proof. Proof is trivial, easily derived from the definition. □

Great. We defined what a convex set is using two vectors. What's in for us if we want more than two vectors. Best we can do for the time being is a *convex combination* of vectors. It's really nothing fancy:

Definition 2 (Convex combination of vectors)**.** *Let there be n vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ and n non-negative coefficients λ_i such that $\lambda_1 + \lambda_2 + \ldots \lambda_n = 1$, any vector $\boldsymbol{z}$ for which the following holds is called a* ***convex combination*** *of the former vectors:*

$$\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 + \ldots + \lambda_n \boldsymbol{x}_n$$

Geometric Interpretation. The set of the convex combinations of n vectors $\boldsymbol{x}_i$ is the *convex hull* of the set of points defined by $\boldsymbol{x}_i$

Theorem 2. A set is convex *iff* it contains all the convex combinations of its elements.

1.3 Preliminaries: Geometric concepts in $\mathbb{R}^d$

We will now move to define some extensions to certain concepts one finds quite familiar and useful in 2-D or 3-D space in order to not only establish a common ground with regards to vocabulary, but also help motivate intuitive geometric thinking as it can prove to be quite valuable and fruitful in the context of convex analysis.

Definition 3. *(Half-space) Given the space $\mathbb{R}^n$, a vector a and a vector b with $a, b \in \mathbb{R}^n$ then the set that is defined by the inequality:*

$$\langle \boldsymbol{a}, \boldsymbol{x} \rangle \leq \boldsymbol{b}, \boldsymbol{x} \in \mathbb{R}^n$$

is called a (closed) half-space of $\mathbb{R}^n$ If the inequality holds strictly (i.e. $\langle \boldsymbol{a}, \boldsymbol{x} \rangle \leq \boldsymbol{b}$) the set is called an open half-space of $\mathbb{R}^n$.

Definition 4. *(Hyper-plane) Let vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ be constant and $\boldsymbol{x} \in \mathbb{R}^n$. The set of points that lie on the set H defined by the equation:*

$$\langle \boldsymbol{a}, \boldsymbol{x} \rangle = \boldsymbol{b}$$

is called a hyper-plane.

Observe that a hyper-plane in $\mathbb{R}^n$ has dimension $n - 1$. The concept of the hyper-plane extends what one naturally would call a line in $\mathbb{R}^2$ (i.e. a sub-space of $\mathbb{R}^2$ with a basis of dimension 1) and a plane in $\mathbb{R}^3$ (i.e. the sub-space of $\mathbb{R}^3$ that has a basis of dimension 2). One more observation is that the vector $\boldsymbol{a}$ is perpendicular to the hyper-plane. Abusing the analogy we draw between liners $\mathbb{R}^2$ and hyper-planes in $\mathbb{R}^d$, we could also refer to $\boldsymbol{a}$ as the *slope* of the hyper-plane.

Definition 5. *(Supporting half-space) Let C be a convex set $C \in \mathbb{R}^n$. A supporting half-space is a half-space that:*

- *contains C*
- *has a point of C on its boundary*

It follows naturally that one would define a *supporting hyper-plane* in the following way:

Definition 6. *(Supporting hyper-plane) The boundary of a supporting half-space of a convex set C is called a supporting half-space.*

1.4 Convex functions

Since we have already defined some fundamental ideas that will function as building blocks for more complex concepts, we will move on to discuss the matters that will mainly concern us, namely convexity with respect to a function.

1.4.1 Definitions

Let f be a function mapping values from $\mathbb{R}^d$ to $\mathbb{R}$. We can imagine f as defining a hyper-surface in the joint space of its input space and its output space, $\mathbb{R}^d \times \mathbb{R}$. The points above that surface whose perpendicular projections on $\mathbb{R}^d$ remain in $\mathrm{dom}f$ form the epigraph of the given function. More formally:

Definition 7. *An* ***epigraph*** *of a fuction $f : \mathbb{R}^n \to \mathbb{R}$ is said to be the set of points $(\boldsymbol{x}, \mu)$ such that $\mu \geq f(\boldsymbol{x})$ and it is noted as:*

$$\mathrm{epi}f = \{(\boldsymbol{x}, \mu) \,|\, \mu \geq f(\boldsymbol{x})\}$$

In 1.1 the epigraphs of two different function can be seen.

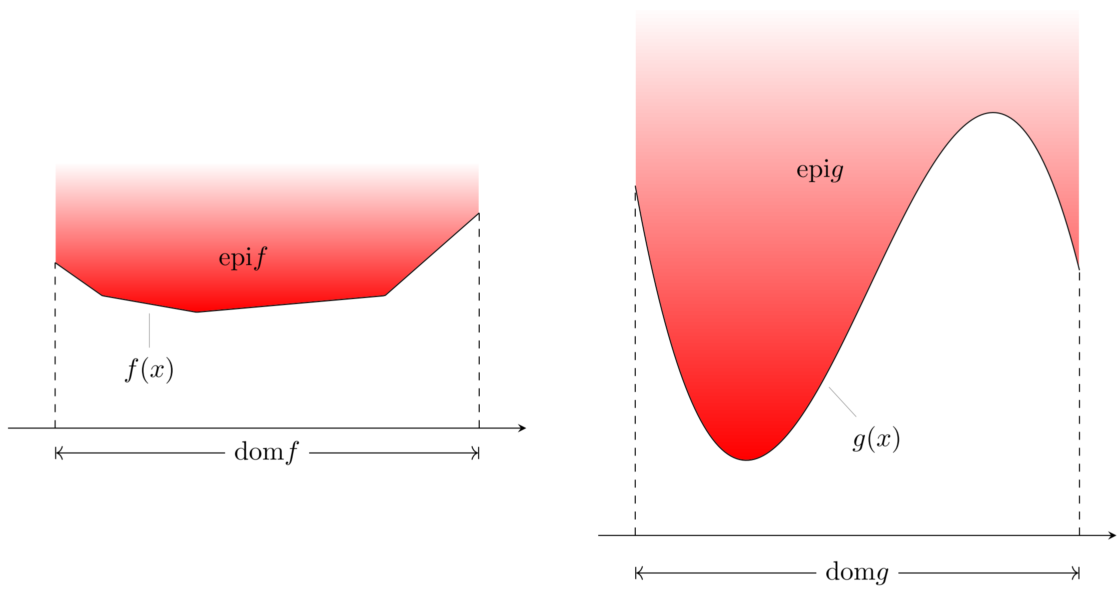

Figure 1.1: Two different functions with their respective epigraphs (the fading red areas)

Definition 8 (Convex function)**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex when* $\mathrm{dom}f$ *is convex and for any $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom}f$ for any $t \in [0, 1]$ the following inequality holds:*

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y})$$

When the latter inequality is strict we say that f is strictly convex.

This definition can extend to what we will call Jensen's inequality, generalizing the inequality from the convex combination of two points to a convex combination of n points something that can at times prove quite useful.

Theorem 3 (Jensen's inequality)**.** Let f be a convex function, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \in \mathrm{dom}f$ and $\lambda_1, \lambda_2, \ldots, \lambda_m \in [0, 1]$ such that $\lambda_1 + \lambda_2 + \ldots + \lambda_m = 1$. Then, it always holds that:

$$f(\lambda_1\boldsymbol{x}_1 + \lambda_2\boldsymbol{x}_2 + \ldots + \lambda_m\boldsymbol{x}_m) \leq \lambda_1 f(\boldsymbol{x}_1) + \lambda_2 f(\boldsymbol{x}_2) + \ldots + \lambda_m f(\boldsymbol{x}_m)$$

Another definition of a convex function is the one that uses a function's epigraph.

Definition 9 (Convex function – alternative definition)**.** *A function is convex when* $\mathrm{epi}f$ *is a convex set.*

We can also demonstrate that the two latter definitions equivalent.

Proof. **(Definition 8 $\Rightarrow$ Definition 9)** Let $(\boldsymbol{x}_1, v_1), \ldots, (\boldsymbol{x}_n, v_n) \in \text{epi}f$. Since the domain of f ($\text{dom}f$) is convex, we know that if we define any set of coefficients $t_1, \ldots, t_n \in [0,1]$ such that $\sum_i t_i = 1$, the point $(\boldsymbol{x}', v') = \left(\sum_i t_i \boldsymbol{x}_i, \sum_i t_i v_i\right)$ will also belong to $\text{dom}f$. By definition of the epigraph, we know that:

$$\begin{cases} v_i \geq f(\boldsymbol{x}_i), \forall i \in \{1, \ldots, n\} \\ v' \geq f(\boldsymbol{x}') \end{cases} \Rightarrow \tag{1.1}$$

$$v' = t_1 v_1 + \ldots + t_n v_n \geq t_1 f(\boldsymbol{x}_1) + \ldots + t_n f(\boldsymbol{x}_n) = \sum_i t_i f(\boldsymbol{x}_i) \tag{1.2}$$

By Jensen's inequality – directly implied by definition 8 – the following will be true:

$$\sum_i t_i f(\boldsymbol{x}_i) \geq f\big(\sum_i t_i \boldsymbol{x}_i\big) \Rightarrow \tag{1.3}$$

$$\sum_i t_i f(\boldsymbol{x}_i) \geq f(\boldsymbol{x}') \tag{1.4}$$

Hence, from 1.2 and 1.4 we see that:

$$v' \geq f(\boldsymbol{x}') \tag{1.5}$$

This means that every point $(\boldsymbol{x}', v')$ defined as the convex combination of points $(\boldsymbol{x}_i, v_i) \in \text{epi}f$, $i = 1, 2, \ldots, n$ belongs in $\text{epi}f$, rendering the latter a convex set.

(Definition 8 $\Leftarrow$ Definition 9)
If $\text{epi}f$ is convex, then for every pair $(\boldsymbol{x}, u)$ and $(\boldsymbol{y}, v)$, for all $t \in [0,1]$, the points $\big(t\boldsymbol{x}(1-t), tu + (1-t)v\big)$ belong in $\text{epi}f$ as well. Since the line interpolated between $(\boldsymbol{x}, u)$ and $(\boldsymbol{y}, v)$ belongs to $\text{epi}f$ every value $f(t\boldsymbol{x} + (1-t)\boldsymbol{y})$ need be less or equal to $tu + (1-t)v$ by definition of the epigraph, that is:

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tu + (1-t)v \tag{1.6}$$

No constraint prevents us from assigning to u, v the values $u = f(\boldsymbol{x})$ and $v = f(\boldsymbol{y})$. Then, 1.6 becomes:

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq t f(\boldsymbol{x}) + (1-t) f(\boldsymbol{y})$$

Since one proposition implies the other, we can decide that they are equivalent. ☐

For the sake of completeness, we will need to give the definitions of some more concepts, namely that of the effective domain, which we have already used without giving a proper definition and that of a proper function.

Definition 10 (Effective Domain of a Convex Function)**.** *The effective domain of a convex function* $\text{dom}f$ *is the set of* $\boldsymbol{x}$ *s.t:*

$$\text{dom}f = \{\boldsymbol{x} \,|\, f(\boldsymbol{x}) < -\infty\}$$

Definition 11 (Proper function)**.** *A function f is called proper if its epigraph is non-empty and contains no vertical lines.*

Proposition 1. Let f be a convex function, f is proper iff there exists at least one point $\boldsymbol{x}$ such that $f(\boldsymbol{x}) < +\infty$ and $f(\boldsymbol{x}) > -\infty$ anywhere else. Or equivalently, its effective domain $\text{dom}f$ is non-empty and f takes at least one finite value.

One more theorem we are going to state about all convex functions, but not yet prove, is the following one that can make us think of convex functions in an intriguing way.

Theorem 4. Every closed convex function f is the pointwise supremum of the colleciton of all affine functions h such that $h \leq f$.

Geometric Interpretation. The latter tells us that if we consider a function $f : \mathbb{R} \to \mathbb{R}$, the curve defined by f can be described at any one of its points as the maximum value of all the lines h for which $h \leq f$.

1.4.2 First and Second Order Conditions

Of course, however easy it might be to grasp the definition, it could fall short in usefulness or practicality with respect to trying to characterize a function as being convex or not. Luckily the next two theorems –restricted on differentiable and twice differentiable functions– can offer a way that can prove quite helpful in characterizing functions as convex or not.

Theorem 5 (First Order Condition)**.** Let a function $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. Then, f is convex iff $\mathrm{dom} f$ is a convex set and for any $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$ the next inequality holds:

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T (\boldsymbol{y} - \boldsymbol{x})$$

Geometric Interpretation. What the latter means is quite simple to comprehend. Given two points $x, y \in \mathrm{dom} f$, regardless of their relative poisition (i.e. it could very well be $x < y$ or $x > y$), if we were to start following the tangent line to the curve that passes trough $\big(x, f(x)\big)$ with horizontal direction that would lead to y, we will consistently find ourselves *below* $f(y)$. We tried to illustrate this in figure 1.2.

We will now prove that under the prerequisite that the function f is differentiable, the the theorem is equivalent to the definition of a convex functions:

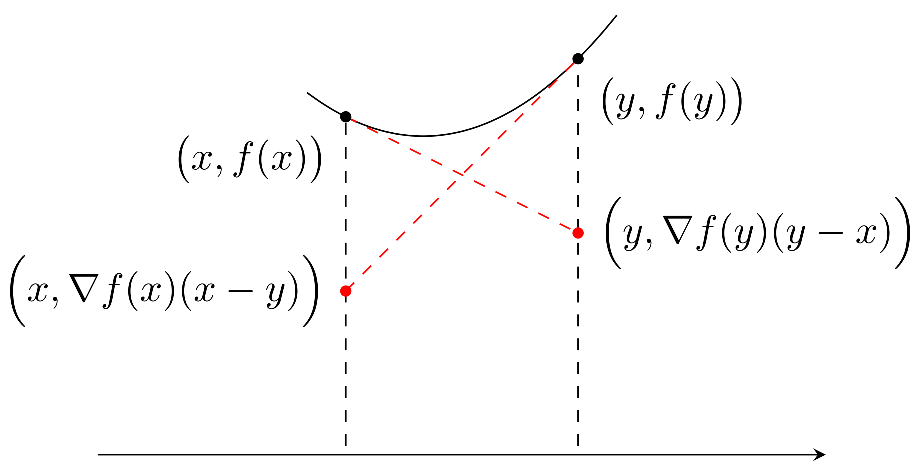

Figure 1.2: Illustration of the first order condition

Proof. We will prove the equivalency with a clever trick; firstly we will prove the claim for functions $\mathbb{R} \to \mathbb{R}$ and the we will parametrize any function $\mathbb{R}^d \to \mathbb{R}$ in a way that will demonstrate the validity of our claim for such functions f with $\mathrm{dom} f \subseteq \mathbb{R}^d$.
(Definition 8 $\Rightarrow$ Theorem 5)

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y) \Rightarrow \tag{1.7}$$
$$f((1-t)x + ty) - f(x) \leq -tf(x) + tf(y) \Rightarrow \tag{1.8}$$
$$f((1-t)x + ty) - f(x) \leq t\big(f(y) - f(x)\big) \xRightarrow{t>0} \tag{1.9}$$
$$\frac{f((1-t)x + ty) - f(x)}{t} \leq f(y) - f(x) \Rightarrow \tag{1.10}$$
$$\frac{f(x + t(y-x)) - f(x)}{t} \leq f(y) - f(x) \Rightarrow \tag{1.11}$$
$$\frac{f(x + t(y-x)) - f(x)}{t(y-x)}(y-x) \leq f(y) - f(x) \xRightarrow{t\to 0} \tag{1.12}$$
$$f'(x)(y-x) \leq f(y) - f(x) \Leftrightarrow \tag{1.13}$$

$$f'(x)(y-x)+f(x)\le f(y) \tag{1.14}$$

(**Definition 8** $\Leftarrow$ **Theorem 5**) For the needs of proving this claim we will consider three points, $x, y, z \in \mathbb{R}$ such that $z = tx + (1-t)y$ which means that $x \le z \le y$. From the theorem in question we would get:

$$\begin{cases} f(x) \ge f(z) + f'(z)(x-z) \\ f(y) \ge f(z) + f'(z)(y-z) \end{cases} \xRightarrow[\times(1-t),0<t<1]{\times t} \tag{1.15}$$

$$\begin{cases} tf(x) \ge tf(z) + tf'(z)(x-z) \\ (1-t)f(y) \ge (1-t)f(z) + (1-t)f'(z)(y-z) \end{cases} \xRightarrow{(+)} \tag{1.16}$$

$$\begin{aligned}
tf(x)+(1-t)f(y) &\ge f(z)+tf'(z)(x-z)+(1-t)f(z)+(1-t)f'(z)(y-z) \Rightarrow && (1.17)\\
&\ge f(z)+tf'(z)x-tf'(z)z+f'(z)y-f'(z)z-tf'(z)y+f'(z)z \Rightarrow && (1.18)\\
&\ge f(z)+tf'(z)x+f'(z)y-f'(z)z-f'(z)y \Rightarrow && (1.19)\\
&\ge f(z)+f'(z)\underbrace{\big(tx+(1-t)y\big)}_{z}-f'(z)z \Rightarrow && (1.20)\\
&\ge f(z)+f'(z)z-f'(z)z \Rightarrow && (1.21)\\
&\ge f(tx+(1-t)y) && (1.22)
\end{aligned}$$

And now we will simply consider a function $f : \mathbb{R}^d \to \mathbb{R}$, two points $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$ and a function $g : \mathbb{R} \to \mathbb{R}$ such that $g(t) = f(t\boldsymbol{x} + (1-t)\boldsymbol{y})$. We derive g with respect to t:

$$\begin{aligned}
g'(t) &= \nabla f(t\boldsymbol{x}+(1-t)\boldsymbol{y})^T \frac{d}{dt}(t\boldsymbol{x}+(1-t)\boldsymbol{y}) \Rightarrow && (1.23)\\
&= \nabla f(t\boldsymbol{x}+(1-t)\boldsymbol{y})^T(\boldsymbol{x}-\boldsymbol{y}) \qquad \text{(this remains a scalar)} && (1.24)
\end{aligned}$$

If we consider f to be convex we can see that this renders g convex as well. Consider two points $\boldsymbol{\sigma}_1 = \tau_1\boldsymbol{x} + (1-\tau_1)\boldsymbol{y}$ and $\boldsymbol{\sigma}_2 = \tau_2\boldsymbol{x} + (1-\tau_2)\boldsymbol{y}$ with $\tau_1, \tau_2 \in [0,1]$. Thanks to f's convexity:

$$\begin{aligned}
f(\lambda\boldsymbol{\sigma}_1+(1-\lambda)\boldsymbol{\sigma}_2) \le \lambda f(\boldsymbol{\sigma}_1)+(1-\lambda)f(\boldsymbol{\sigma}_2) \Leftrightarrow && (1.25)\\
g(\lambda\tau_1+(1-\lambda)\tau_2) \le \lambda g(\tau_1)+(1-\lambda)g(\tau_2) && (1.26)
\end{aligned}$$

Convrsely, if g is convex then for $\tau_1, \tau_2 \in [0,1]$:

$$g(\lambda\tau_1+(1-\lambda)\tau_2) \le \lambda g(\tau_1)+(1-\lambda)g(\tau_2) \tag{1.27}$$

If we choose $\tau_1 = 0$ and $\tau_2 = 1$:

$$\begin{aligned}
g(\lambda) \le \lambda g(0)+(1-\lambda)g(1) \Leftrightarrow && (1.28)\\
f(\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{y}) \le \lambda f(\boldsymbol{x})+(1-\lambda)f(\boldsymbol{y}) && (1.29)
\end{aligned}$$

We successfully showed that for differentiable functions with a domain that is a subset of $\mathbb{R}$, theorem 5 is equivalent to definition 8 and by means of the simple device of function g we were able to generalize the equivalency claim to function with a domain in $\mathbb{R}^d$. □

Theorem 6 (Second Order Condtion)**.** Let a function f be twice differentiable and $\mathrm{dom} f$ convex. Function f is convex iff for every $\boldsymbol{x} \in \mathrm{dom} f$:

$$\nabla^2 f(\boldsymbol{x}) \succeq 0$$

(i.e. the Hessian matrix[1]of f is positive semi-definite or $\boldsymbol{z}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{z} \ge 0, \forall \boldsymbol{z}$)

[1]Reminder: $\nabla^2 = \begin{pmatrix} \frac{\partial^2}{\partial x_1^2} & \frac{\partial^2}{\partial x_1\,\partial x_2} & \cdots & \frac{\partial^2}{\partial x_1\,\partial x_n} \\ \frac{\partial^2}{\partial x_2\,\partial x_1} & \frac{\partial^2}{\partial x_2^2} & \cdots & \frac{\partial^2}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n\,\partial x_1} & \frac{\partial^2}{\partial x_n\,\partial x_2} & \cdots & \frac{\partial^2}{\partial x_n^2} \end{pmatrix}$

Proof. The rationale for proving the second order condition is similar to proving the first order condition. First, we prove that the equivalence stands for functions with a single dimensional domain and then extend the claim to prove it to domains with more than one dimensions.

For a function f with $\text{dom}f$, theorem 6 translates to $f''(x) \geq 0$. It should be well established that this means $f'(x)$ is non-decreasing.

Let points $x, y, z \in \text{dom}f$ with $x \leq z \leq y$ and $z = tx + (1-t)y$ for some $t \in [0,1]$. Then we can conclude that since $f'(x)$ is non-decreasing:

$$f'(x) \leq f'(z) \leq f'(y)$$

It is obvious that f is an antiderivative of f', hence:

$$\begin{cases} \int_x^z f'(t)dt = f(z) - f(x) \\ \int_z^y f'(t)dt = f(y) - f(x) \end{cases} \tag{1.30}$$

By utilizing the Mean Value Theorem for integrals from calculus, we can decide that there exist points ξ_1, ξ_2 for which $x \leq \xi_1 \leq z \leq \xi_2 \leq y$ such that:

$$\begin{cases} \int_x^z f'(t)dt = f(\xi_1)(z-x) \\ \int_z^y f'(t)dt = f(\xi_2)(y-z) \end{cases} \xrightarrow[z\leq\xi_2]{z\geq\xi_1} \tag{1.31}$$

$$\begin{cases} \int_x^z f'(t)dt \leq f'(z)(z-x) \\ \int_z^y f'(t)dt \geq f'(z)(y-z) \end{cases} \Rightarrow \tag{1.32}$$

$$\begin{cases} f(z) - f(x) \leq f'(z)(z-x) \\ f(y) - f(z) \geq f'(z)(y-z) \end{cases} \xrightarrow{z=tx+(1-t)y} \tag{1.33}$$

$$\begin{cases} f(z) - f(x) \leq f'(z)\big(tx + (1-t)y - x\big) \\ f(y) - f(z) \geq f'(z)\big(y - tx - (1-t)y\big) \end{cases} \Rightarrow \tag{1.34}$$

$$\begin{cases} f(z) - f(x) \leq f'(z)(1-t)(y-x) \\ f(y) - f(z) \geq f'(z)t(y-x) \end{cases} \xrightarrow[\times(1-t)]{\times t} \tag{1.35}$$

$$\begin{cases} tf(z) - tf(x) \leq tf'(z)(1-t)(y-x) \\ (1-t)f(y) - (1-t)f(z) \geq (1-t)f'(z)t(y-x) \end{cases} \Rightarrow \tag{1.36}$$

$$\begin{cases} tf(z) - tf(x) \leq tf'(z)(1-t)(y-x) \\ (1-t)f(z) - (1-t)f(y) \leq (1-t)f'(z)t(x-y) \end{cases} \xrightarrow{(+)} \tag{1.37}$$

$$f(z) - tf(x) - (1-t)f(y) \leq 0 \Rightarrow \tag{1.38}$$

$$f(z) \leq tf(x) + (1-t)f(y) \xrightarrow{z=tx+(1-t)y} \tag{1.39}$$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \qquad \text{(convex by definition)} \tag{1.40}$$

As for the converse, we shall show that by reductio ad absurdum the second derivative of a convex function cannot receive negative values. We it to receive a negative value, then by it being continuous (it is differentiable), there should by an interval $[q, r] \subseteq \text{dom}f$ such that inequalities would take the opposite direction:

$$\begin{cases} \int_x^z f'(t)dt > f'(z)(z-x) \\ \int_z^y f'(t)dt < f'(z)(y-z) \end{cases} \tag{1.41}$$

This would inevitably lead to the following inequality if we were to follow the same steps as we just previously did:

$$f(tx + (1-t)y) > tf(x) + (1-t)f(y)$$

But, we considered f to be convex, hence if f is convex there cannot be any interval of its domain where its second derivatives should receive negative values.

As for the case of the multidimensional domain. We consider g such that $g(t) = f(t\boldsymbol{x} + (1-t)\boldsymbol{y})$. As we showed while proving theorem 5, g is convex iff f is convex. We also proved that g is convex iff $g''(t) \geq 0$. If we try to derive g twice we would get the following:

$$g''(t) = \frac{d^2}{dt^2} g(t) \Rightarrow \tag{1.42}$$

$$= \frac{d^2}{dt^2} f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \Rightarrow \tag{1.43}$$

$$= \frac{d}{dt}\Big(\nabla f(t\boldsymbol{x} + (1-t)\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})\Big) \Rightarrow \tag{1.44}$$

$$= (\boldsymbol{x} - \boldsymbol{y})^T \nabla^2 f(t\boldsymbol{x} + (1-t)\boldsymbol{y})(\boldsymbol{x} - \boldsymbol{y}) \tag{1.45}$$

We can observe that if $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$ by choosing an appropriate t, the vector $(\boldsymbol{x} - \boldsymbol{x})$ could be any vector in $\mathbb{R}^d$. Hence, for g to be convex, the Hessian of f needs to be positive semi-definite. □

1.5 Conjugate Transform & Fenchel's inequality

In this section we are going to discuss the conjugate transform of functions. It is a transform that maps the parameters of hyper-planes tangent to the curve of a function to a certain value. It may not be the first time one sees such a transform, one that shifts our attention to a parameter space. For example, in traditional computer vision a certain transform, known as Hough Transform, is used in order to map whole lines of the 2-D space to tuples (ρ, θ); θ being the angle that the line perpendicular to the line in question forms with the horizontal axis and ρ being the distance of the line from the origin. Although this only vaguely resembles the subject of our discussion – and we regret causing any confusion – we mention it in order to motivate more ways of thinking of lines than just as a set of points. There are various implementations based on this premise that help us detect and recognize not only lines but also regular geometric shapes.

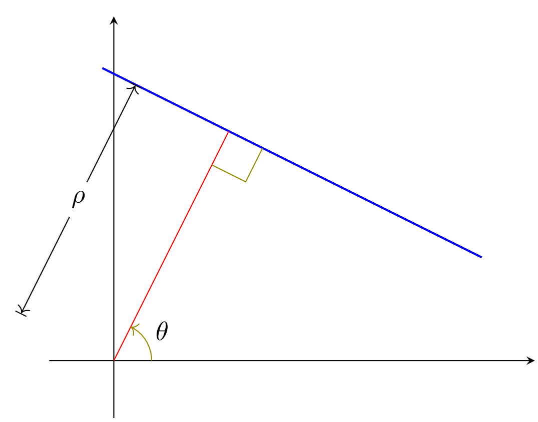

Figure 1.3: Representing the blue line with parameters ρ, θ

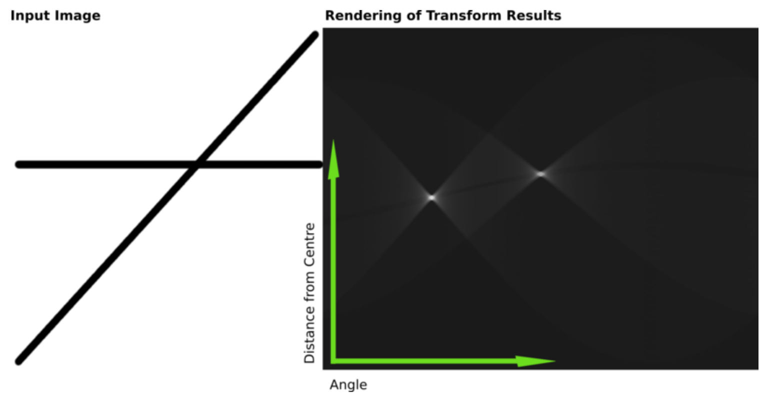

Figure 1.4: An implementation of the Hough Transform

Our subject revolves around tangent lines (or hyper-planes for function domains with dimension greater than 1) on a convex function. We will demonstrate a way that has been devised in order to represent elegantly the whole set of these tangent lines.

Definition 12 (Convex conjugate)**.** *Let a function f be convex. We define its conjugate transform as the function f^* such that:*

$$f^*(\boldsymbol{p}) = \sup_{\boldsymbol{x} \in \mathrm{dom} f} \{\langle \boldsymbol{p}, \boldsymbol{x} \rangle - f(\boldsymbol{x})\}$$

Frankly the definition seems a bit awkward. Considering its geometric interpretation could maybe shed some light as to what this is supposed to mean.

Geometric Interpretation. The conjugate transform of a function f is merely a function f^* that maps slopes α to the maximum available offset β such that the given line $\alpha x + \beta$ will be tangent to the curve defined by f.

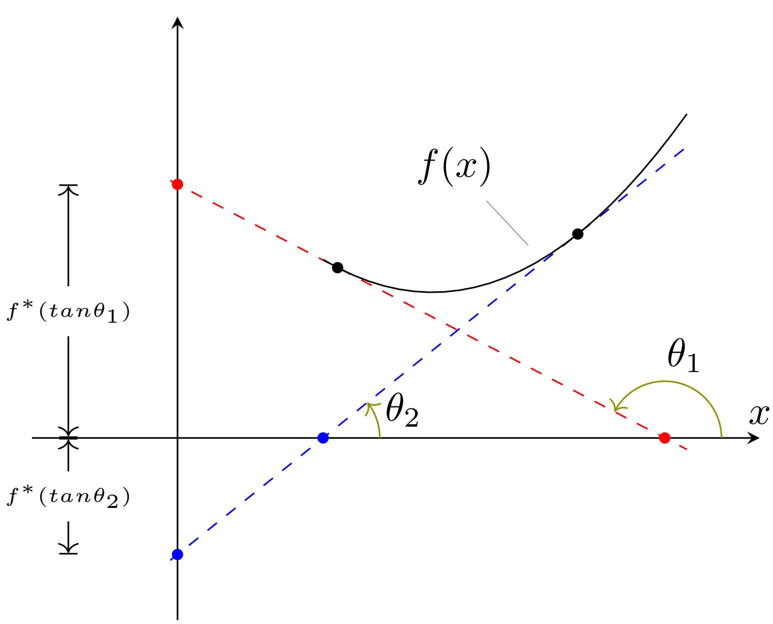

Figure 1.5: Geometric meaning of the conjugate transform

The conjugate transform f^* of a function, f if certain conditions hold for the latter, can give us all the information we need about f. Keeping in mind the theorem about f being described as the point-wise supremum of all affine functions h such that $h \leq f$, it seems rather intuitive. We will state this formally:

Theorem 7 (Fenchel-Moreau Theorem)**.** Let f be lower semi-continuous[2] and convex, then:

$$f^{**} = f$$

1.6 Subgradients, subdifferentials

1.6.1 Definition

This section is concerned with function that are not everywhere differentiable. Although we cannot define a gradient at a given point, it may be sufficient to substitute an exact gradient with slope-vectors that will always undeshoot the value of function in question for any given pair of points.

However informal is the initial description of the subgradient might have been, the formal definition does not fall far:

Definition 13. *(Subgradient) Let a function f be convex. Any vector $\boldsymbol{p}$ is called a subgradient of f at a point $\boldsymbol{x} \in \mathrm{dom} f$ if for any $\boldsymbol{z} \in \mathrm{dom} f$:*

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{p}, \boldsymbol{z} - \boldsymbol{x} \rangle$$

Definition 14. *(Subdifferential) The set of every subgradient vector of a convex function f at point $\boldsymbol{x}$ is called the subdifferential, $\partial f(\boldsymbol{x})$, of f at point $\boldsymbol{x}$.*

In figure 1.6 we tried to illustrate a number of subgradients for the function f that is not everywhere differentiable.

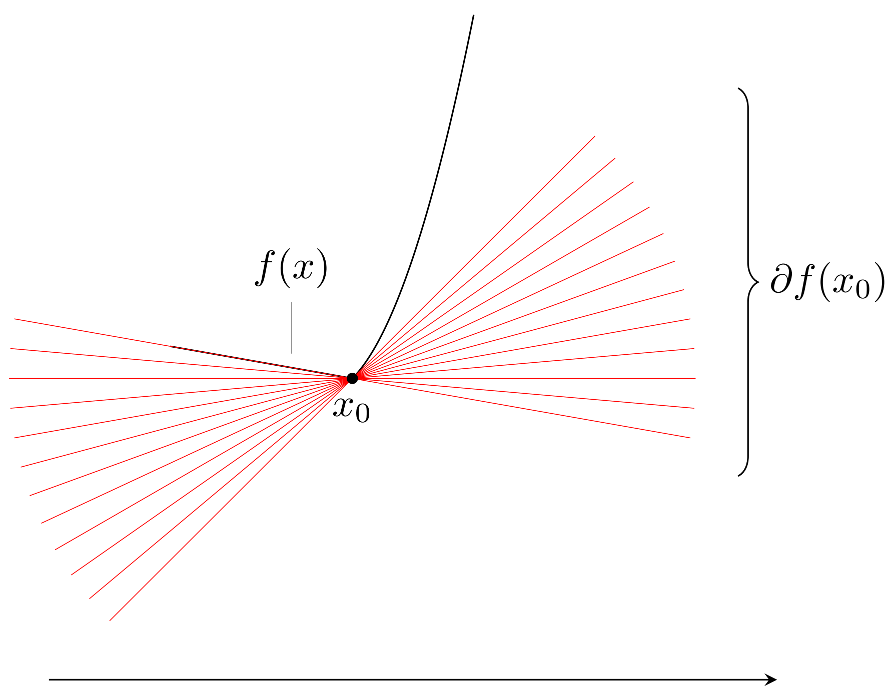

Figure 1.6: The subdifferential $\partial f(x_0)$ is the set of all subgradients which are represented by the red lines. A subgradient at point x_0 will always undershoot the value of f at any point z.

Since we have seen a quite intuitive definition of the subdifferential, why not use our intuition as a stepping stone in order to grasp a more technical definition of it? Before we move on to define subgradients alternatively, we will need to remind ourselves the notion of directional derivatives.

[2]Reminder: A function is lower semi-continuous at x_0 if for every $\epsilon > 0$ there exists a neighborhood U of x_0 such that $f(x) \geq f(x_0) - \epsilon \; \forall x \in U$ if $f(x_0) < +\infty$. Else if $f(x_0) \to +\infty$ then $f(x) \to +\infty$ as well.

Definition 15 (Directional Derivative)**.** *The directional derivative $f'(\cdot,\cdot)$ of a function f at point $\boldsymbol{x}$ in direction $\boldsymbol{d}$ is defined as:*

$$f'(\boldsymbol{x},\boldsymbol{d}) = \lim_{t\downarrow 0} \frac{f(\boldsymbol{x}+t\boldsymbol{d}) - f(\boldsymbol{x})}{t}$$

(Obviously, $t > 0$)

We can now define the subdifferential with respect to the directional derivative.

Definition 16 (Subdifferential–Alternative Definition)**.** *The subdifferential ∂f of f at point $\boldsymbol{x}$ is the set defined as such:*

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{s} \,|\, \langle \boldsymbol{s},\boldsymbol{d}\rangle \leq f'(\boldsymbol{x},\boldsymbol{d}),\, \forall \boldsymbol{d} \in \mathbb{R}^d\}$$

Vectors $\boldsymbol{s}$ are the subgradients of f at point $\boldsymbol{x}$.

1.6.2 Fenchel's inequality

Since we have now seen both the conjugate transform and the definition of the subgradient, we can move on to state Fenchel's inequality:

Theorem 8 (Fenchel's inequality)**.** For any subgradient vector $\boldsymbol{p} \in f^*(\mathrm{dom} f^*)$ and any $\boldsymbol{x} \in \mathrm{dom} f$ the following inequality stands:

$$f^*(\boldsymbol{p}) + f(\boldsymbol{x}) \geq \langle \boldsymbol{p},\boldsymbol{x}\rangle$$

Proof. By the definition of the conjugate transform: $f^*(\boldsymbol{p}) = \sup_{\boldsymbol{x}\in\mathrm{dom} f}\{\langle \boldsymbol{p},\boldsymbol{x}\rangle - f(\boldsymbol{x})\}$ we can decide that:

$$f^*(\boldsymbol{p}) \geq \langle \boldsymbol{p},\boldsymbol{x}\rangle - f(\boldsymbol{x}), \forall \boldsymbol{x} \Leftrightarrow \tag{1.46}$$

$$f^*(\boldsymbol{p}) + f(\boldsymbol{x}) \geq \langle \boldsymbol{p},\boldsymbol{x}\rangle \tag{1.47}$$

□

1.6.3 Mean Value Theorems

Subgradients seem to magically lift restrictions placed upon the validity of known theorems when functions are non-differentiable. An instance of this is the reinstatement of the mean value theorem for functions that are convex but not everyhwere differentiable.

Theorem 9 (Differential form)**.** Let a function $f : \mathbb{R}^d \to \mathbb{R}$ be convex. For any $\boldsymbol{x},\boldsymbol{y} \in \mathrm{dom} f$, there will always exist a $t \in (0,1)$ for which $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$ and $\boldsymbol{s} \in \partial f(\boldsymbol{z})$ such that:

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \langle \boldsymbol{z}, \boldsymbol{y}-\boldsymbol{x}\rangle$$

Theorem 10 (Integral form)**.** Let a function $f : \mathbb{R}^d \to \mathbb{R}$ be convex. For any $\boldsymbol{x},\boldsymbol{y} \in \mathrm{dom} f$ with $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$ and any collection of subgradients on points $\boldsymbol{z}(t)$, $\forall t \in [0,1]$ (i.e. for every point between $\boldsymbol{x}$ and $\boldsymbol{y}$):

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = \int_{t=0}^{1} \langle \partial f(\boldsymbol{z}(t)), \boldsymbol{y}-\boldsymbol{x}\rangle \, dt$$

1.6.4 Extending convexity theorems to non-differentiable functions

We can now extend some previously stated theorems for the convexity of differentiable functions to functions that are not everywhere differentiable.

Lemma 1. The following are equivalent:

(a) f is convex

(b) (First Order Condition Analogue) $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{s}_x^T(\boldsymbol{y}-\boldsymbol{x})$

(c) (Monotonicity of Subgradients) $(\boldsymbol{s}_y - \boldsymbol{s}_x)^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0$, $\forall \boldsymbol{x}, \boldsymbol{y}$ and any $\boldsymbol{s}_x \in \partial f(\boldsymbol{x})$, $\boldsymbol{s}_y \in \partial f(\boldsymbol{y})$

Proof. The path from (a) to (b) is rather simple. We write down the definition of convexity and through the directional derivative reach to the subgradient:

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) \Rightarrow \tag{1.50}$$
$$f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) \Rightarrow \tag{1.51}$$
$$f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - f(\boldsymbol{y}) \leq tf(\boldsymbol{x}) - tf(\boldsymbol{y}) \Rightarrow \tag{1.52}$$
$$\frac{f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - f(\boldsymbol{y})}{t} \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) \xRightarrow{t\downarrow 0} \tag{1.53}$$
$$f'(\boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y}) \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) \xRightarrow{\text{by definition}} \tag{1.54}$$
$$\langle \boldsymbol{s}_y, \boldsymbol{y} - \boldsymbol{x} \rangle \leq f'(\boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y}) \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) \tag{1.55}$$

□

As for the path from (b) to (a), we need two pairs of points $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{z}, \boldsymbol{y}$ – with $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$ – to apply (b) on:

$$\begin{cases} f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \boldsymbol{s}_z^T(\boldsymbol{x} - \boldsymbol{z}) \\ f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \boldsymbol{s}_z^T(\boldsymbol{y} - \boldsymbol{z}) \end{cases} \xRightarrow{\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}} \tag{1.56}$$
$$\begin{cases} f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + (1-t)\boldsymbol{s}_z^T(\boldsymbol{x} - \boldsymbol{y}) \\ f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + t\boldsymbol{s}_z^T(\boldsymbol{y} - \boldsymbol{x}) \end{cases} \xRightarrow[\times(1-t)]{\times t} \tag{1.57}$$
$$\begin{cases} tf(\boldsymbol{x}) \geq tf(\boldsymbol{z}) + t(1-t)\boldsymbol{s}_z^T(\boldsymbol{x} - \boldsymbol{y}) \\ (1-t)f(\boldsymbol{y}) \geq (1-t)f(\boldsymbol{z}) + (1-t)t\boldsymbol{s}_z^T(\boldsymbol{y} - \boldsymbol{x}) \end{cases} \xRightarrow{(+)} \tag{1.58}$$
$$t(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) \geq f(\boldsymbol{z}) \Rightarrow \tag{1.59}$$
$$t(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) \geq f(t\boldsymbol{X} + (1-t)\boldsymbol{y}) \tag{1.60}$$

Hence, (a)⇔(b).

With regards to (c), we can observe that applying (b) for both swaps of $\boldsymbol{x}, \boldsymbol{y}$ and by adding by parts we get:

$$\begin{cases} f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{s}_x^T(\boldsymbol{y} - \boldsymbol{x}) \\ f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \boldsymbol{s}_y^T(\boldsymbol{x} - \boldsymbol{y}) \end{cases} \xRightarrow{(+)} \tag{1.61}$$
$$f(\boldsymbol{y}) + f(\boldsymbol{x}) \geq f(\boldsymbol{x}) + f(\boldsymbol{y}) + \boldsymbol{s}_x^T\boldsymbol{y} - \boldsymbol{x} + \boldsymbol{s}_y^T(\boldsymbol{x} - \boldsymbol{y}) \Rightarrow \tag{1.62}$$
$$(\boldsymbol{s}_y - \boldsymbol{s}_x)^T(\boldsymbol{y} - \boldsymbol{x}) \geq 0 \tag{1.63}$$

As for reaching (b) from (c), we have to revert to the integral form of the mean value theorem for subgradients after stating (c) for $\boldsymbol{y}$ and an intermediate point $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$:

$$(\boldsymbol{s}_z - \boldsymbol{s}_y)^T(\boldsymbol{z} - \boldsymbol{x}) \geq 0 \Rightarrow \tag{1.64}$$
$$(\boldsymbol{s}_z - \boldsymbol{s}_y)^T\left(t\boldsymbol{x} + (1-t)\boldsymbol{y} - \boldsymbol{x}\right) \geq \Rightarrow \tag{1.65}$$
$$t(\boldsymbol{s}_z - \boldsymbol{s}_y)^T(\boldsymbol{x} - \boldsymbol{y}) \geq 0 \Rightarrow \tag{1.66}$$
$$(\boldsymbol{s}_z - \boldsymbol{s}_y)^T(\boldsymbol{x} - \boldsymbol{y}) \geq 0 \Rightarrow \tag{1.67}$$
$$\boldsymbol{s}_z^T(\boldsymbol{x} - \boldsymbol{y}) \geq \boldsymbol{s}_y^T(\boldsymbol{x} - \boldsymbol{y}) \tag{1.68}$$

Which means, the regular derivative of $g(t)$ is the directional derivative of $f(t\boldsymbol{x} + (1-t)\boldsymbol{y})$ on the direction $\boldsymbol{x} - \boldsymbol{y}$, since we will have to restrict it to right side assuming $t > 0$. Now, from the mean value theorem of differential calculus we get:

$$g(1) - g(0) = f(\boldsymbol{x}) - f(\boldsymbol{y}) = \int_0^1 \left(f'(t\boldsymbol{x} + (1-t)\boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y})\right) dt \tag{1.69}$$

We now know that since inequality 1.68 holds[3]:

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) = \int_0^1 \big(f'(t\boldsymbol{x} + (1-t)\boldsymbol{y}, \boldsymbol{x} - \boldsymbol{y})\big)\, dt \tag{1.70}$$
$$= \int_0^1 \big(f'(\boldsymbol{z}, \boldsymbol{x} - \boldsymbol{y})\big)\, dt \Rightarrow \tag{1.71}$$
$$\geq \int_0^1 \big(\boldsymbol{s}_z^T(\boldsymbol{y} - \boldsymbol{x})\big)\, dt \xRightarrow{1.68} \tag{1.72}$$
$$\geq \int_0^1 \big(\boldsymbol{s}_x^T(\boldsymbol{y} - \boldsymbol{x})\big)\, dt \Rightarrow \tag{1.73}$$
$$= \big(\boldsymbol{s}_x^T(\boldsymbol{y} - \boldsymbol{x})\big) \int_0^1 dt \Rightarrow \tag{1.74}$$
$$= \boldsymbol{s}_x^T(\boldsymbol{y} - \boldsymbol{x}) \Rightarrow \tag{1.75}$$
$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \geq \boldsymbol{s}_x^T(\boldsymbol{y} - \boldsymbol{x}) \tag{1.76}$$

1.7 Strong Convexity

As we can observe, if a function is convex it means that we can bound every value of the function if it lies between two other given values of that function. There is one more – quite stronger – bound we can define for certain functions. The latter functions are the ones that we are going to call *strongly convex*. In strongly convex functions, values of the function intermediate to any pair of points $\boldsymbol{x}, \boldsymbol{y} \in \mathrm{dom} f$ can be bound with combination of the previously stated bound and an appropriate parabola (more or less what we would call a bowl).

Since there is not much consensus in the most popular texts on the matter and although some authors choose to speak of strong convexity in the context of twice differentiable, we chose to give a more liberal definition of strong convexity –following Hiriart-Urruty and Lemaréchal (2001)– that is not immediately concerned with differentiability.

Definition 17. *(Strong Convexity) A function $f : \mathbb{R}^m \to \mathbb{R}$ is said to be a strongly convex function with coefficient (more precisely, modulus) m if the latter inequality holds:*

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{m}{2}t(1-t)\|x - y\|_2^2$$

is convex.

Lemma 2. A function f is strongly convex if $f(\boldsymbol{x}) - \frac{m}{2}\|\boldsymbol{x}\|^2$ is convex.

Remark 1. If $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, then f is strongly convex iff the next sentence holds for some $m > 0$:

$$\nabla^2 f \succeq mI$$

Proof. We need to observe that the latter implies:

$$\nabla^2 f(\boldsymbol{x}) - mI \succeq 0 \Leftrightarrow \tag{1.77}$$
$$\nabla^2\Big(f(\boldsymbol{x}) - \frac{m}{2}\boldsymbol{x}^T\boldsymbol{x}\Big) \succeq 0 \tag{1.78}$$

This means that the function $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{m}{2}\boldsymbol{x}^T\boldsymbol{x}$ [4] needs to be convex. As we very well know, by means of the second order condition the claim of equivalency must hold since if we write down the inequality stated by definition 8 we will derive the mentioned inequality:

$$g(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tg(\boldsymbol{x}) + (1-t)g(\boldsymbol{x}) \Rightarrow \tag{1.79}$$

[3] Reminder:

$$f(x) \geq g(x) \Rightarrow \int_a^b f(x)\, dx \geq \int_a^b g(x)\, dx$$

[4] Reminder: $\boldsymbol{x}^T\boldsymbol{x} = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|_2^2$

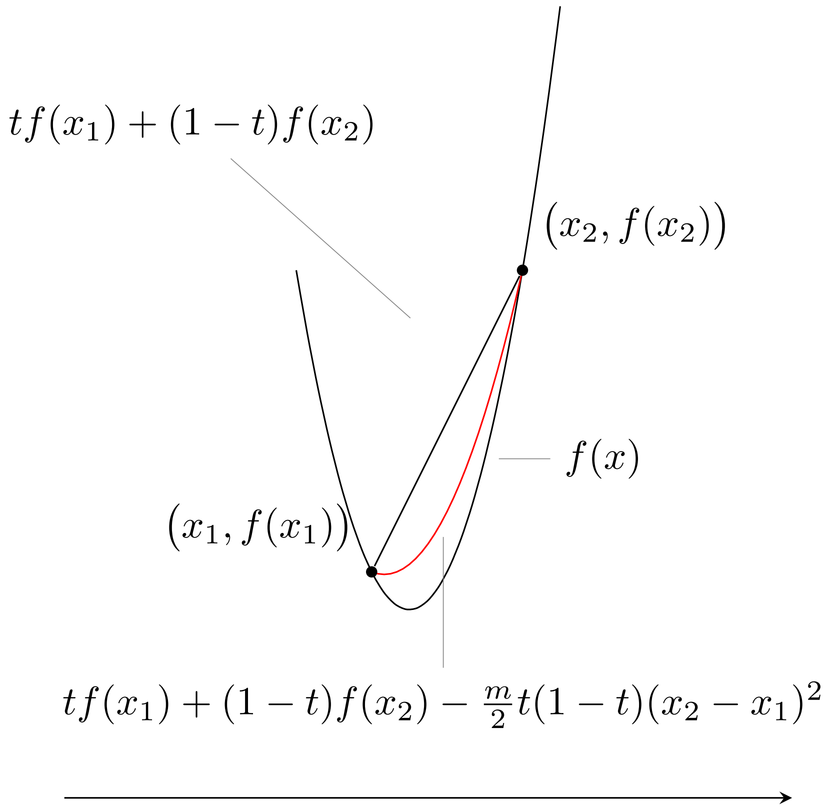

Figure 1.7: A strongly convex function is not only bounded by the straight black line but the red line as well

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) - m(t\boldsymbol{x} + (1-t)\boldsymbol{y})^T(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq t\Big(f(\boldsymbol{x}) - m\boldsymbol{x}^T\boldsymbol{x}\Big) + (1-t)\Big(f(\boldsymbol{y}) - m\boldsymbol{y}^T\boldsymbol{y}\Big) \Rightarrow \tag{1.80}$$

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{m}{2}t(1-t)(\boldsymbol{y} - \boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) \Leftrightarrow \tag{1.81}$$

$$f(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{m}{2}t(1-t)\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \tag{1.82}$$

□

Lemma 3. The following sentences are equivalent to f being strongly convex with modulus m (or m-strongly-convex):

(a) $$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{s}^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \tag{1.83}$$

(b) $$(\boldsymbol{s}_y - \boldsymbol{s}_x)^T(\boldsymbol{y} - \boldsymbol{x}) \geq m\|\boldsymbol{y} - \boldsymbol{x}\|_2^2 \tag{1.84}$$

Proof. In order to get from (a) to definition 17 we apply it on two pairs of points, $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{z}, \boldsymbol{y}$, where $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y})$, accordingly. We will then multiply both instances with $(t-1)$ and t accordingly and sum by parts to get the desired inequality:

$$\begin{cases} f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \boldsymbol{s}_z^T(\boldsymbol{y} - \boldsymbol{z}) + \frac{m}{2}\|\boldsymbol{y} - \boldsymbol{z}\|^2 \\ f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \boldsymbol{s}_z^T(\boldsymbol{x} - \boldsymbol{z}) + \frac{m}{2}\|\boldsymbol{x} - \boldsymbol{z}\|^2 \end{cases} \Rightarrow \tag{1.85}$$

$$\begin{cases} f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \boldsymbol{s}_z^T\big(\boldsymbol{y} - (t\boldsymbol{x} + (1-t)\boldsymbol{y})\big) + \frac{m}{2}\|\boldsymbol{y} - (t\boldsymbol{x} + (1-t)\boldsymbol{y})\|^2 \\ f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + \boldsymbol{s}_z^T\big(\boldsymbol{x} - (t\boldsymbol{x} + (1-t)\boldsymbol{y})\big) + \frac{m}{2}\|\boldsymbol{x} - (t\boldsymbol{x} + (1-t)\boldsymbol{y})\|^2 \end{cases} \Rightarrow \tag{1.86}$$

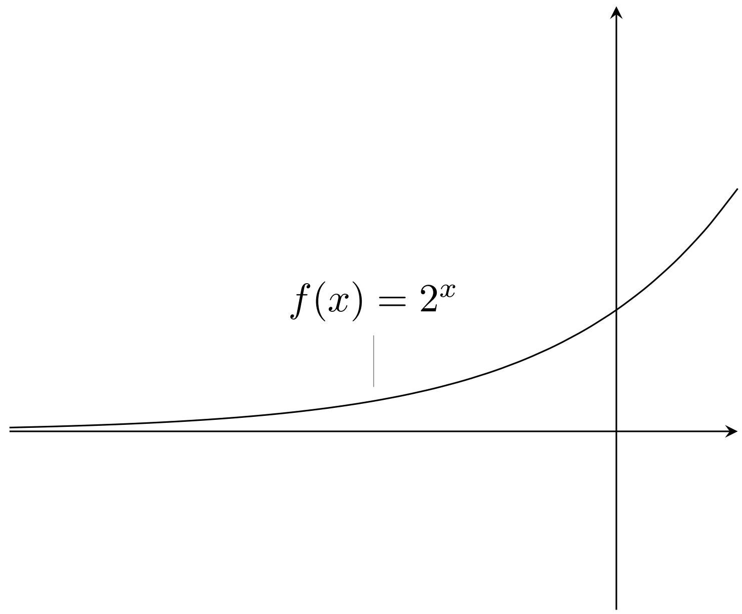

Figure 1.8: Function $f(x) = 2^x$ is not strongly convex as $f''(x)$ can be arbitrarily close to 0 for $x < 0$

$$\begin{cases} f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + t\boldsymbol{s}_z^T(\boldsymbol{y}-\boldsymbol{x}) + t^2\frac{m}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \\ f(\boldsymbol{x}) \geq f(\boldsymbol{z}) + (1-t)\boldsymbol{s}_z^T(\boldsymbol{x}-\boldsymbol{y}) + (1-t)^2\frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \end{cases} \xRightarrow[\times t]{\times(t-1)} \tag{1.87}$$

$$\begin{cases} (t-1)f(\boldsymbol{y}) \geq (t-1)\Big(f(\boldsymbol{z}) + t\boldsymbol{s}_z^T(\boldsymbol{x}-\boldsymbol{y}) + t^2\frac{m}{2}\|\boldsymbol{x}-\boldsymbol{z}\|^2\Big) \\ tf(\boldsymbol{x}) \geq t\Big(f(\boldsymbol{z}) + (1-t)\boldsymbol{s}_z^T(\boldsymbol{y}-\boldsymbol{x}) + (1-t)^2\frac{m}{2}\|\boldsymbol{y}-\boldsymbol{z}\|^2\Big) \end{cases} \xRightarrow{(+)} \tag{1.88}$$

$$tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) \geq f(\boldsymbol{z}) + \frac{m}{2}t(t-1)\|y-x\|^2 \tag{1.89}$$

Conversely, can use the definition 17 to get to (a) by using the alternative definition of the subdifferential that encompasses the directional derivative:

$$f(t\boldsymbol{x}+(1-t)\boldsymbol{y}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{m}{2}t(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \xRightarrow{\boldsymbol{z}=t\boldsymbol{x}+(1-t)\boldsymbol{y}} \tag{1.90}$$

$$f(\boldsymbol{z}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{m}{2}t(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \Rightarrow \tag{1.91}$$

$$f(\boldsymbol{z}) \leq tf(\boldsymbol{x}) + (1-t)f(\boldsymbol{y}) - \frac{m}{2}t(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \Rightarrow \tag{1.92}$$

$$f(\boldsymbol{z}) - f(\boldsymbol{y}) \leq tf(\boldsymbol{x}) - tf(\boldsymbol{y}) - \frac{m}{2}t(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \Rightarrow \tag{1.93}$$

$$\frac{f(\boldsymbol{z}) - f(\boldsymbol{y})}{t} \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) - \frac{m}{2}(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \Rightarrow \tag{1.94}$$

$$\frac{f(\boldsymbol{z}) - f(\boldsymbol{y})}{t} + \frac{m}{2}(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) \Rightarrow \tag{1.95}$$

$$\frac{f(\boldsymbol{y}+t(\boldsymbol{x}-\boldsymbol{y})) - f(\boldsymbol{y})}{t} + \frac{m}{2}(1-t)\|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) \xRightarrow{t\downarrow 0} \tag{1.96}$$

$$f'(\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) \xRightarrow{\text{def.}16} \tag{1.97}$$

$$\langle \boldsymbol{s}_y, \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq f'(\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{y}) \tag{1.98}$$

The latter is just the definition of strong convexity with $\boldsymbol{x}, \boldsymbol{y}$ swapped.

Now, we can move from (a) to (b) as such:

$$\begin{cases} f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{s}_x^T(\boldsymbol{y}-\boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \\ f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \boldsymbol{s}_y^T(\boldsymbol{x}-\boldsymbol{y}) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \end{cases} \xRightarrow{(+)} \tag{1.99}$$

$$f(\boldsymbol{y}) + f(\boldsymbol{x}) \geq f(\boldsymbol{y}) + \boldsymbol{s}_y^T(\boldsymbol{x}-\boldsymbol{y}) + \frac{m}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2 \boldsymbol{s}_x^T(\boldsymbol{y}-\boldsymbol{x}) + f(\boldsymbol{x}) + \frac{m}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \Rightarrow \tag{1.100}$$

$$0 \geq \boldsymbol{s}_y^T(\boldsymbol{x}-\boldsymbol{y}) + \boldsymbol{s}_x^T(\boldsymbol{y}-\boldsymbol{x}) + m\|\boldsymbol{y}-\boldsymbol{x}\|^2 \Rightarrow \tag{1.101}$$

$$-\boldsymbol{s}_y^T(\boldsymbol{x}-\boldsymbol{y}) - \boldsymbol{s}_x^T(\boldsymbol{y}-\boldsymbol{x}) \geq m\|\boldsymbol{y}-\boldsymbol{x}\|^2 \Rightarrow \tag{1.102}$$

$$\boldsymbol{s}_y^T(\boldsymbol{y}-\boldsymbol{x}) - \boldsymbol{s}_x^T(\boldsymbol{y}-\boldsymbol{x}) \geq m\|\boldsymbol{y}-\boldsymbol{x}\|^2 \Rightarrow \tag{1.103}$$

$$(\boldsymbol{s}_y^T - \boldsymbol{s}_x^T)(\boldsymbol{y}-\boldsymbol{x}) \geq m\|\boldsymbol{y}-\boldsymbol{x}\|^2 \Rightarrow \tag{1.104}$$

$$(\boldsymbol{s}_y - \boldsymbol{s}_x)^T(\boldsymbol{y}-\boldsymbol{x}) \geq m\|\boldsymbol{y}-\boldsymbol{x}\|^2 \tag{1.105}$$

And now we can close the circle of arguments, using (b) to return to (a). We start from (b) for points $\boldsymbol{y}$ and $\boldsymbol{z} = t\boldsymbol{x} + (1-t)\boldsymbol{y}$ and we get:

$$\langle \boldsymbol{s}_z - \boldsymbol{s}_y, \boldsymbol{z}-\boldsymbol{y}\rangle \geq m\|\boldsymbol{z}-\boldsymbol{y}\|^2 \Rightarrow \tag{1.106}$$

$$\langle \boldsymbol{s}_z - \boldsymbol{s}_y, \big(t\boldsymbol{x} + (1-t)\boldsymbol{y} - \boldsymbol{y}\big)\rangle \geq m\|\boldsymbol{y} - \big(t\boldsymbol{x} + (1-t)\boldsymbol{y} - \boldsymbol{y}\big)\|^2 \Rightarrow \tag{1.107}$$

$$\langle \boldsymbol{s}_z - \boldsymbol{s}_y, t(\boldsymbol{x}-\boldsymbol{y})\rangle \geq m\|t(\boldsymbol{x}-\boldsymbol{y})\|^2 \Rightarrow \tag{1.108}$$

$$t\langle \boldsymbol{s}_z - \boldsymbol{s}_y, \boldsymbol{x}-\boldsymbol{y}\rangle \geq mt^2\|(\boldsymbol{x}-\boldsymbol{y})\|^2 \Rightarrow \tag{1.109}$$

$$\langle \boldsymbol{s}_z - \boldsymbol{s}_y, \boldsymbol{x}-\boldsymbol{y}\rangle \geq mt\|(\boldsymbol{x}-\boldsymbol{y})\|^2 \Rightarrow \tag{1.110}$$

$$\langle \boldsymbol{s}_z, \boldsymbol{x}-\boldsymbol{y}\rangle \geq \langle \boldsymbol{s}_y, \boldsymbol{x}-\boldsymbol{y}\rangle + mt\|(\boldsymbol{x}-\boldsymbol{y})\|^2 \tag{1.111}$$

We revert to the function $g(t) = f(t\boldsymbol{x} + (1-t)\boldsymbol{y})$ which is convex as we have seen. We apply the integral form of the mean value theorem for subgradients to get the following relation which we will combine (b) with 1.111 :

$$g(1) - g(0) = f(\boldsymbol{x}) - f(\boldsymbol{y}) = \int_0^1 \langle \boldsymbol{s}_z, \boldsymbol{x}-\boldsymbol{y}\rangle\, dt \xRightarrow{1.111} \tag{1.112}$$

$$\geq \int_0^1 \Big(\langle \boldsymbol{s}_y, \boldsymbol{x}-\boldsymbol{y}\rangle + mt\|(\boldsymbol{y}-\boldsymbol{x})\|^2\Big)\, dt \Rightarrow \tag{1.113}$$

$$= \int_0^1 \Big(\langle \boldsymbol{s}_x, \boldsymbol{x}-\boldsymbol{y}\rangle\Big)\, dt + \int_0^1 t\, dt\, m\|(\boldsymbol{x}-\boldsymbol{y})\|^2 \Rightarrow \tag{1.114}$$

$$= \langle \boldsymbol{s}_x, \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{1}{2}m\|(\boldsymbol{x}-\boldsymbol{y})\|^2 \Rightarrow \tag{1.115}$$

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) \geq \langle \boldsymbol{s}_y, \boldsymbol{x}-\boldsymbol{y}\rangle + \frac{m}{2}\|(\boldsymbol{x}-\boldsymbol{y})\|^2 \tag{1.116}$$

□

1.8 Lipschitz Condition

One way to implicitly speak of functions whose (sub-)gradients are bounded is to use the definition of them being λ-Lipschitz or not, specifically when the slope of every one of its tangents is bounded by λ. Such a function is said to satisfy the Lipschitz Condition.

Definition 18. *(Lipschitz Condition) Let a function $\boldsymbol{f}: \mathbb{R}^m \to \mathbb{R}^n$. The function is said to be λ-Lipschitz or to satisfy a Lipschitz Condition if there exists a constant $\lambda \in \mathbb{R}, \lambda > 0$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$:*

$$\|\boldsymbol{f}(\boldsymbol{y}) - \boldsymbol{f}(\boldsymbol{x})\| \leq \lambda\|\boldsymbol{y}-\boldsymbol{x}\|$$

We complement the definition with two examples of one function that satisfies the Lipschitz Condition and is not everywhere continuous and one that is continuous everywhere and does satififsy the Lipschitz Condition.

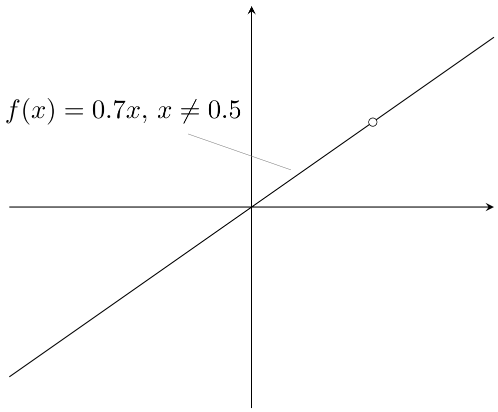

Figure 1.9: The straight line that has one point of incontinuity satisfies the Lipschitz Condition

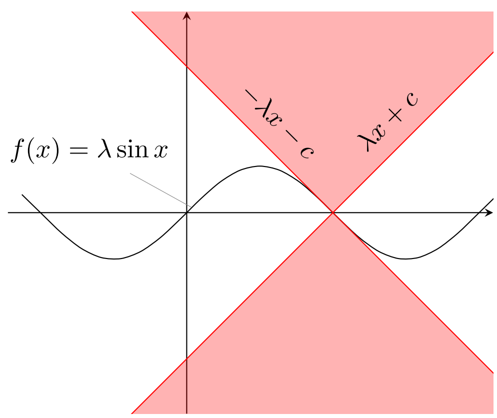

Figure 1.10: The function $\lambda \sin x$ has tangents that at any given point, never get into the red area defined by the two red lines for different c's.

References

Boyd, S., Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2001). *Fundamentals of Convex Analysis*. Springer Berlin Heidelberg.
URL http://link.springer.com/10.1007/978-3-642-56468-0

Rockafellar, R. T. (1997). *Convex analysis*, vol. 36. Princeton university press.