Convex Optimization by Prof. Stephen Becker

Jaden Wang

February 8, 2021

Contents

Chapter 1

Theoretical Foundation

1.1 Introduction

An optimization problem looks like

$$\min_{x \in C} f(x)$$

where $f(x)$ is the **objective function** and $C \subseteq \mathbb{R}^n$ is the **constraint set**. C might look like

$$C = \{x : g_i(x) \leq 0 \ \forall \ i = 1, \ldots, m\}.$$

Remark. We can always turn a maximization problem into a minimization problem as the following:

$$\min_x f(x) = -\max_x -f(x).$$

Therefore, WLOG, we will stick with minimization.

Example. An assistant professor earns \$100 per day, and they enjoy both ice cream and cake. The optimization problem aims to maximize the utility (*e.g.* happiness) of ice cream $f_1(x_1)$ and of cake $f_2(x_2)$. The constraints we have is that $x_1 \geq 0, x_2 \geq 0$, and $x_1 + x_2 \leq 100$.

To maximize both utility, it might be natural to define

$$F(\operatorname{vec} x) = \begin{pmatrix} f_1(x_1) \\ f_2(x_2) \end{pmatrix}, \operatorname{vec} x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

and maximize F. However, this isn't a well-defined problem, because *there is no total order on* $\mathbb{R}^n$! That is, we don't have a good way to compare whether a vector is bigger than another vector, except in the cases when the same direction of inequality can be achieved for all components of two vectors and a partial order can be established. For this kind of **multi-objective** optimization problem, we can look for Pareto-optimal points in these special cases. We can also try to convert the output into a scalar as the following:

$$\min_x f_1(x) + \lambda \cdot f_2(x_2)$$

for some $\lambda > 0$ that reflects our preference for cake vs ice cream. But this can be subjective.

Thus, For the remainder of this class, we are only going to assume $f : \mathbb{R}^n \to \mathbb{R}$.

Moreover, for $f : \mathbb{R} \to \mathbb{R}$, it's very easy to solve by using root finding algorithms or grid search. So since interesting problems occur with vector inputs, we will simply use x to represent vectors.

Notation. min asks for the minimum value, whereas arg min asks for the minimizer that yields the minimum value.

1.1.1 Lipschitz continuity

Example. Let's consider a variant of the Dirichlet function, $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \begin{cases} x & \text{if } x \in \mathbb{Q} \\ 1 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

Then the solution to the problem

$$\min_{x \in [0,1]} f(x) = 0$$

is $x = 0$ by observation. However, the function is not smooth and a small perturbation can yield wildly different values. Thus, it is not tractable to solve this numerically.

This requires us to add a smoothness assumption:

Definition: Lipschitz continuity

$f : \mathbb{R}^n \to \mathbb{R}$ is ***L*-Lipschitz continuous** with respect to a norm $\|\cdot\|$ if for all $x, y \in \mathbb{R}^n$,

$$|f(x) - f(y)| \leq L \cdot \|x - y\|.$$

Note. Lipschitz continuity implies continuity and uniform continuity. It is a stronger statement because it tells us *how* the function is (uniformly) continuous. However, it doesn't require differentiability.

Definition: l_p norms

For $1 \leq p < \infty$,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

For $p = \infty$,

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|.$$

Remark. $\|x\|_1$ and $\|x\|_2^2$ have separable terms as they are sums of their components. $\|x\|_2^2$ is also differentiable which makes it the nicest norm to optimize.

Example. Let $f : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz continuous w.r.t. $\|\cdot\|_\infty$. Let $C = [0,1]^n$, *i.e.* in $\mathbb{R}^2$, C is a square. To solve the problem

$$\min_{x \in C} f(x),$$

since we have few assumption, there is no better method (in the worst case sense) than the **uniform grid method**. The idea is that we pick $p+1$ points in each dimension, *i.e.* $\{0, \frac{1}{p}, \frac{2}{p}, \dots, 1\}$, so we would have $(p+1)^n$ points in total.

Let x^* be a global optimal point, then there exists a grid point $\widetilde{x}$ s.t.

$$\|x^* - \widetilde{x}\|_\infty \le \frac{1}{2} \cdot \frac{1}{p}.$$

Thus by Lipschitz continuity,

$$\begin{aligned} |f(x^*) - f(\widetilde{x})| &\le L \cdot \|x^* - \widetilde{x}\|_\infty \\ &\le \frac{1}{2}\frac{L}{p} \end{aligned}$$

So we can find $\widetilde{x}$ by taking the discrete minimum of all $(p+1)^n$ grid points.

In (non-discrete) optimization, we usually can't exactly find the minimizer, but rather find something very close.

Definition: epsilon-optimal solution

x is a **ε-optimal solution** to $\min_{x \in C} f(x)$ if $x \in C$ and

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

where $f^* = \min_{x \in C} f(x)$.

Our uniform grid method gives us an ε-optimal solution with $\varepsilon = \frac{L}{2p}$, and requires $(p+1)^n$ function evaluations. Writing p in terms of ε, we have $p = \frac{L}{2\varepsilon}$

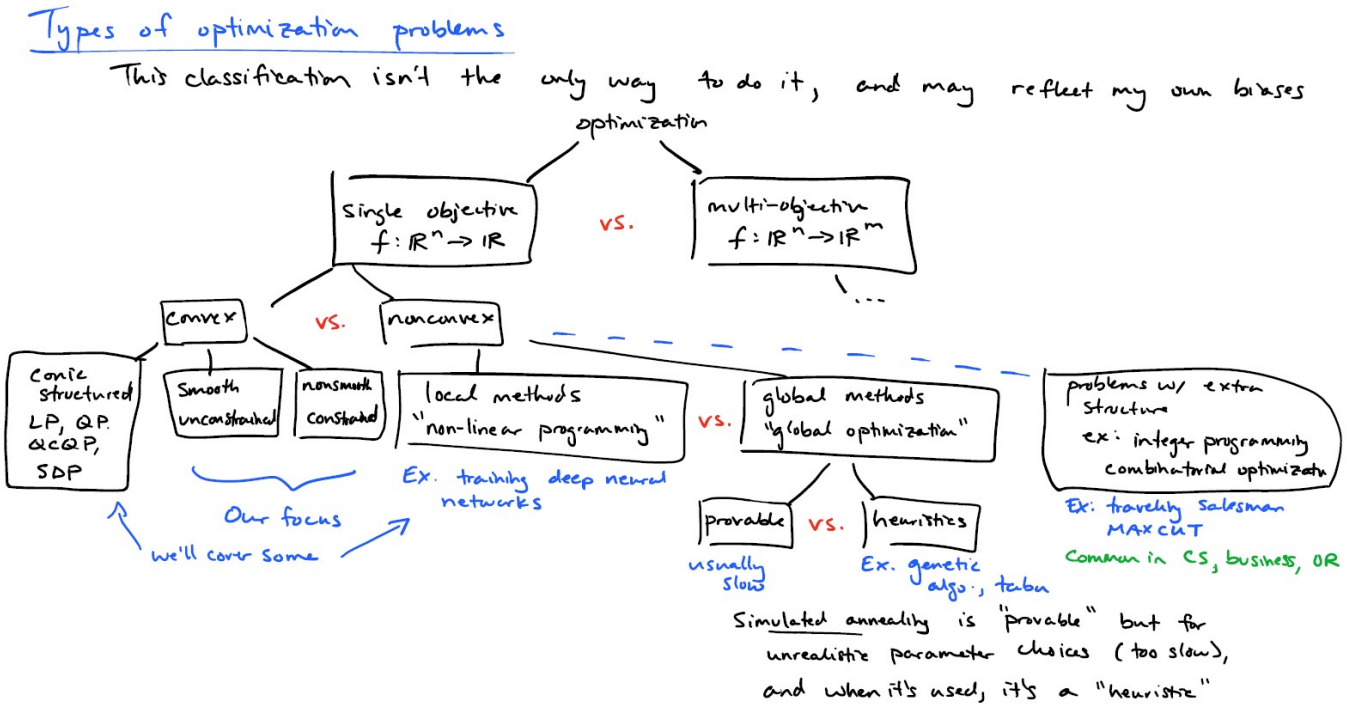
so equivalently it requires $\left(\frac{2L}{\varepsilon}+1\right)^n$ function evaluations, which approximately is ε^{-n}.

For $\varepsilon = 10^{-6}$, $n = 100$, it requires 10^{600} function evaluations. This is really bad!

Take-aways from this example:

- curse-of-dimensionality: there can be trillions of variables in a Google Neural Network. It would be intractable using the grid method.
- we need more assumptions to allow us to use more powerful methods.

1.1.2 Categorization



1.1.3 Minimizers

We are given a generic problem $\min_{x\in C} f(x), C \subseteq \mathbb{R}^n$. Then a **feasible point** x means $x \in C$. A **solution** or **minimizer** or **global minimizer** x^* means

1) $x^* \in C$

2) $\forall\ y \in C, f(x^*) \leq f(y)$

It might not be unique, *i.e.* $x^* \in \arg\min_{x\in C} f(x)$.

Example.

$$\min_{x\in\mathbb{R}} f(x) \text{ where } f(x) = 0\ \forall\ x.$$

Sometimes the solution may not exist (even for convex problems).

Example.

$$\min_{x\in(0,1)} x^2.$$

x^* is a **local minimizer** if x^* is feasible and there exists an $\varepsilon > 0$ s.t. $f(x^*) \leq f(y)\quad \forall\ y \in C \cap B_\varepsilon(x^*) \coloneqq \{y : \|y - x^*\|\} < \varepsilon$. A **strict local minimizer** simply doesn't achieve equality. x^* is an **isolated local minimum** if it is a local minimum and no other local minimum are nearby. Notice that isolated implies strict but the converse is false.

Example (strict but not isolated)**.**

$$f(x) = \begin{cases} x^4 \cos\left(\frac{1}{x}\right) + 2x^4 & x \neq 0 \\ 0 & x = 0 \end{cases}$$

$x^* = 0$ is strict but not isolated due to the rapid oscillation near $x = 0$.

Notation. $f \in \mathcal{C}^3$ means f, f', f'', f''' all exist and are continuous. $f \in \mathcal{C}^3(\mathbb{R}^n)$ means $f, \nabla f, \nabla^2 f, \nabla^3 f$ all exists and are continuous.

Connections with Calculus 1

Recall that in Cal 1, we first find the stationary/critical points in the domain. Then we add the boundary points and minimize over the small (finite) set of candidates.

In high-dimension optimization, we cannot check critical points and the boundary separately because the set of points in the boundary becomes infinite. Moreover, there can be infinite critical points too.

Necessary condition: if x^* is a local or global minimizer and $C = \mathbb{R}^n$, then x^* is a **critical point**. But the converse is false.

Notation. The boundary of C is denoted as $\partial C \coloneqq \overline{C} \setminus \operatorname{int} C$.

If x^* is a critical point but is not a local or global minimizer, then it's a **saddle point**.

Theorem: Weierstrass

If f is continuous and C is compact, then f achieves its infimum over C. That is,

$$\inf_{x \in C} f(x) = \min_{x \in C} f(x).$$

Note. This is pretty much the same as the Extreme Value Theorem.

Proof

First let's prove a claim.

Claim. Every compact set K is closed and bounded.

Closed: suppose not, the compact set K doesn't contain all its limit points. That is, there exists a limit point $x \notin K$ s.t. a sequence $(x_n) \subseteq K$ converges to x. But that also means that all subsequences of (x_n) converges to $x \notin K$ as well, contradicting with the definition of compactness that for every sequence in K there exists a subsequence that converges inside K.

Bounded: suppose not, for all $M > 0$, there exists a $x \in K$ s.t. $\|x\| > M$. This allows us to find a sequence $(x_n) \subseteq K$ s.t. $x_n > n$. This way every subsequence is also unbounded and cannot converge, contradicting with the definition of sequential compactness.

Now let's begin proof proper. Since C is compact and f is continuous, the image of C under f, $f(C)$, is also compact (this follows from sequen-

tial definition of continuity). By the claim $f(C)$ is bounded and closed, meaning that it has an infimum (completeness axiom) and contains the infimum (closed). Thus, f achieves its infimum over C. □

Remark. It would be nice if our constraints C are compact. But at the very least, we want our constraint sets to be closed. For example, $\|Ax - b\| \leq \varepsilon$ instead of $\|Ax - b\| < \varepsilon$.

Several things to note about the feasible set C:

If $C = \emptyset$, the problem is infeasible. This is not always easy to spot.

In this class, C will usually be convex and not integral, *i.e.* $\mathbb{Z}^n$.

Integral constraint is problematic because the optimal integer solution might not be at all close to the optimal real solution, so we cannot obtain it by solving for the real solution first and then round it.

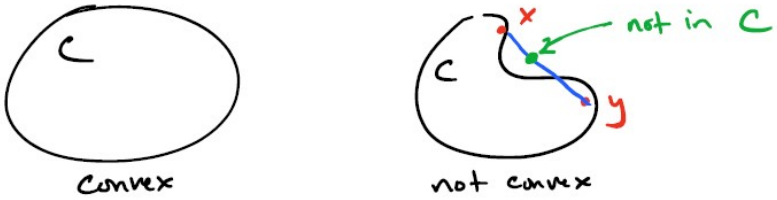
1.1.4 Convexity

Note. From now on we always assume the constraint set C is a subset of a vector space.

Definition: convex set

A set C is **convex** if for all $x, y \in C$ and for all $t \in [0, 1]$, then

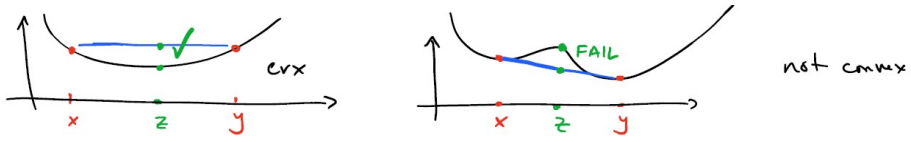
$$tx + (1 - t)y \in C.$$



Definition: convex function

$f : \mathbb{R}^n \to \mathbb{R}$ is a **convex function** if for all $x, y \in \mathbb{R}^n$ and $t \in [0, 1]$, then

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$



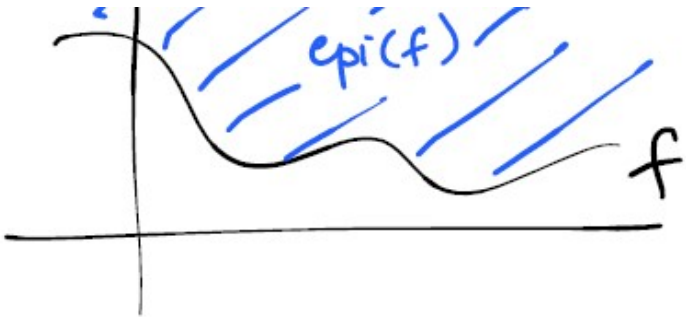
Remark. Linear or affine functions are both convex and concave.

Recall that a **graph** of a function is just the set of points we use to plot a function (now generalized to functions with domain of any dimension).

Definition: epigraph

$$\text{epi}(f) = \{(x, s) : x \in \mathbb{R}^n, s \in \mathbb{R}, s \geq f(x)\}.$$

Intuition. The epigraph of f is sort of the "upper" partition of the vector space that the graph of f resides and partitions. We use epigraph to bridge the concepts of convex sets and convex functions.



Proposition

A function f is convex if and only if the set $\text{epi}(f)$ is convex.

Theorem

If f is convex and C is convex, then any local minimizer of $\min_{x \in C} f(x)$ is in fact global. The set of global solutions is also convex and in particular connected.

This is very neat for optimization!

Proof

Given a local minimizer $x^* \in C$, let's show that it is a global minimizer. Suppose not, that is, there exists a point $x \in C$ s.t. $f(x) < f(x^*)$. Since x^* is a local minimizer, there exists an $\varepsilon > 0$ s.t. $f(x^*) \leq f(y)$ for all $y \in C$ s.t. $\|y - x^*\| < \varepsilon$. Clearly $\|x^* - x\| \geq \varepsilon$ or x^* would not be a local minimizer. Choose $t < \frac{\varepsilon}{\|x^* - x\|} \in [0, 1]$. Since C is convex, we know that the point $x_0 = tx + (1-t)x^* \in C$. Notice that

$$\begin{aligned} \|x^* - x_0\| &= \|x^* - (tx + (1-t)x^*)\| \\ &= \|t(x^* - x)\| \\ &= t\|x^* - x\| \\ &< t \cdot \frac{\varepsilon}{t} \\ &= \varepsilon \end{aligned}$$

That is, x_0 in the ε-neighborhood of x^* and it follows that $f(x^*) \leq f(x_0)$. Since f is convex,

$$\begin{aligned} f(x_0) &= f(tx + (1-t)x^*) \\ &\leq tf(x) + (1-t)f(x^*) \\ &< tf(x^*) + (1-t)f(x^*) \\ &= f(x^*) \end{aligned}$$

This contradicts with the fact that $f(x^*) \leq f(x_0)$. Hence we prove that any local minimizer x^* must also be a global minimizer.

To show that the set of global minimizers is connected, it suffices to prove that it is path-connected. The path we check is of course the line segment in the definition of convex set:

$$g(t) = ta + (1-t)b, \qquad a, b \in \operatorname*{argmin}_{x \in C} f(x).$$

It's easy to see that $f(g(t)) \leq tf(a) + (1-t)f(b) = \min_{x \in C} f(x)$ for all $t \in [0,1]$. It follows that $f(g(t))$ must equal to the global minimum for all $t \in [0,1]$. This makes $g(t) \in \operatorname{argmin}_{x \in C} f(x)\ \forall\ t \in [0,1]$. Thus, the continuous function $g(t)$ is the path we seek. Since a, b are arbitrary global minimizers, we kill two birds in one stone and show that the set of global minimizers is 1. convex and 2. path-connected and therefore connected. □

1.2 Convex Sets

1.2.1 Convex, affine, and cone

Definition

Let $x, y \in \mathbb{R}^n$ (or any vector space), then

1) $tx + (1-t)y, t \in [0,1]$ is a **convex combination** (of x, y).
2) $tx + (1-t)y, t \in \mathbb{R}$ is a **linear combination**.
3) $tx + sy, t, s \geq 0$ is a **(convex) conic combination**.

Definition

A set $C \subseteq \mathbb{R}^n$ is

1) **convex** if for all $x, y \in C$, it contains all convex combinations of x, y.
2) **affine** if for all $x, y \in C$, it contains all linear combinations of x, y.
3) a **cone** if for all $x \in C$, it contains all conic combinations of x.
4) a **convex cone** if it's convex and a cone. That is, for all $x, y \in C, t, x \geq 0$, $tx + sy \in C$.

Note. Affine implies convex based on definition.

Remark. An affine set/subspace is like a subspace except it is possible shifted (may not include 0). Think inhomogenous equation from differential equations. It's also analogous to cosets.

Recall from analysis, the **closure** of A, $\overline{A}$, is the union of A and all its limit points. We can also characterize $\overline{A}$ as the smallest closed set containing A or equivalently the intersection of all closed sets containing A. We can do something similar here.

Definition: affine hull

The **affine hull** of C, $\text{aff}(C)$, is the smallest affine set containing C.

The **affine dimension** of C is $\dim(\text{aff}(C))$. For example, although the unit circle in $\mathbb{R}^2$ has dimension 1, its affine hull is all of $\mathbb{R}^2$ so its affine dimension is 2.

Definition: convex hull

The **convex hull** of C, $\text{conv}(C)$, is the smallest convex set containing C. It is equivalent to the set of all convex combinations of points in C:

$$\left\{\sum_{i=1}^{k} t_i x_i : x_i \in C, t_i \geq 0, \sum_{i=1}^{k} t_i = 1\right\}.$$

Intuition. Given an arbitrary set C, we wrap a rubber band around it and the region enclosed by the rubber band is $\text{conv}(C)$.

Definition: conic hull

The **conic hull** of C is the set of all conic combinations of points in C. That is,

$$\left\{\sum_{i=1}^{k} t_i x_i : x_i \in C, t_i \geq 0\right\}.$$

It is the smallest convex cone that contains C.

Definition: relative interior

The **relative interior** of a set C is the set

$$\text{ri}(C) = \{x \in C : \ \exists\ \varepsilon > 0, B_\varepsilon(x) \cap \text{aff}(C) \subseteq C\}$$

Note. This is really useful for studying symmetric matrices.

Example. Let $C = [0,1] \subseteq \mathbb{R}$, then $\text{int}(C) = \text{ri}(C) = (0,1)$.

However, if $C = [0,1] \times \{0\}$ which has the same shape but is a embedding in $\mathbb{R}^2$, then $\text{int}(C) = \emptyset$ because for every $x \in C$, $B_\varepsilon(x)$ goes outside C along the second dimension. But $\text{ri}(C) = (0,1)$ because $\text{aff}(C) = \mathbb{R} \times \{0\} \cong \mathbb{R}$.

1.2.2 Important examples

Definition: hyperplane

For $a \in \mathbb{R}^n, b \in \mathbb{R}$, the **hyperplane** would be the affine, $n-1$ dimensional set

$$\{x \in \mathbb{R}^n : a^T x = b\}.$$

Alternatively,

$$\{x \in \mathbb{R}^n : a^T (x - x_0) = 0\}.$$

Note. Hyperplanes are convex and affine with $n-1$ dimension. In 2D, it's a line. In 3D it's an actual plane. Also recall from cal 3 that a is the normal vector of the hyperplane.

Definition: half-space

A hyperplane partition $\mathbb{R}^n$ into two **half-spaces**. They have the form

$$\{x \in \mathbb{R}^n : a^T x \leq b\}.$$

Note. Half spaces are convex but not affine.

Definition: Euclidean ball

Open ball: $B_\varepsilon(x) = \{y \in \mathbb{R}^n : \|y - x\| < \varepsilon\}$.

Closed ball: $\overline{B_\varepsilon}(x) = \{y \in \mathbb{R}^n, \|y - x\| \leq \varepsilon\}$.

Note. Balls are convex but not affine.

Definition: ellipsoid

An **ellipsoid** has the form

$$\mathcal{E} = \{x : (x - x_0)^T P^{-1}(x - x_0) \leq 1\}$$

for some matrix $P \succ 0$.

Notation. $A \succ 0$ in this course means A is symmetric and positive definite.

Note. Ellipsoid, like ball, is convex but not affine. If we choose $P = \varepsilon^2 I$, then we get an ε-ball.

Intuition. This is a generalization of Cal 3 ellipsoid using quadratic form. Recall that the P^{-1} in the middle of $x^T x$ is giving us a *weighted* sum. Since $y^T y = \|y\|^2 \leq 1$ is a unit ball, a weighted norm would help us transform an ellipsoid into the unit ball. We use the inverse of P in the definition because the image of this quadratic form is sort of a unit ball, but we are more interested in knowing how to go from the unit ball to the ellipsoid, and P encodes this transformation. Also since $P \succ 0$, using Spectral Theorem we can find the principle axises and length of the ellipsoid.

Example (cones)**.**

- positive orthant in $\mathbb{R}^n$, *e.g.* first quadrant in $\mathbb{R}^2$.

  $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n : x_i \geq 0\}$ is a cone.

  However, $\mathbb{R}^n_{++} = \{x \in \mathbb{R}^n : x_i > 0\}$ is not a cone as a cone must include the additive identity 0 to be closed under non-negative scalar multiplication.

- Lorentz cone/2nd order cone/"ice cream cone"

  $$C = \{(x, t) \in \mathbb{R}^{n+1}, x \in \mathbb{R}^n, t \in \mathbb{R} : \|x\|_2 \leq t\}.$$

- the set of positive semidefinite matrices (PSD): this is the most important nonploy hedral cone. We assume PSDs are Hermitian and are denoted by $A \succeq 0$.

  Notation. $\mathbb{S}^n$ denotes the set of symmetric $n \times n$ matrices. Similar to the reals, we use $\mathbb{S}^n_+$ to denote the set of symmetric positive semidefinite

matrices.

Definition: polyhedron

A **polyhedron** $\mathcal{P} \subseteq \mathbb{R}^n$ is a set of the intersection of a *finite* number of half-spaces and hyperplanes. That is,

$$\mathcal{P} = \{x \in \mathbb{R}^n : a_j^T x \leq b_j, j = 1, \ldots, m; c_j^T x = d_j, j = 1, \ldots, p\}.$$

Note. Intersection of infinite number of half-spaces and hyperplanes are not necessarily a polyhedron (in the intuitive sense) because it can "smooth out" the edges and turn it into for example a ball, such as

$$\overline{B_1}(0) = \{x \in \mathbb{R}^n : a_j^T x \leq 1, a_j \in \text{ unit circle}\}.$$

Note. Polyhedra are always convex since it's finite intersection of convex sets.

Note. The terms "polygon", "polyhedron", and "polytope" will be used interchangeably in this course.

Recall that if a set of points are *linearly independent*, then their linear combinations can equal zero only if all coefficients are zero. Moreover, in an n-dimensional vector space, there can at most be n linearly independent points.

Definition: affinely independent

A set of points $\{x_i\}_{i=0}^n$ is **affinely independent** if

$$\sum_{i=1}^{n} t_i(x_i - x_0) = 0 \Rightarrow t_i = 0.$$

Note. It doesn't matter which x_i we choose to be x_0. They are all equivalent. This is because an affine space can be think of as a translation of a vector space. That is, every element in the affine space is an element from a vector space offset by the same translation vector. When we subtract any two elements from the affine space, the translation vector cancels out and leaves us an element from the vector space so we go back to linear independence in the vector space. Again we can think of the solutions to inhomogeneous differential equations for a concrete example.

Note. In a n-dimensional vector space, at most $n+1$ points can be affinely independent. The "+1" comes from bringing the 0 in the vector space up to x_0, elevating the n-dimensional vector space to an n-dimensional embedding in a $n+1$-dimensional vector space. For example, $\mathbb{R}^n$ requires at least n points to fully describe it via the span. We can visualize $\mathbb{R}^n$ as an origin-containing plane embedded in $\mathbb{R}^{n+1}$. If we translate $\mathbb{R}^n$ by a vector x_0, we get an n-dimensional affine space that now requires at least $n+1$ points to describe.

Definition: simplex

For any set of $k+1$ affinely independent points $\{x_i\}_{i=0}^n$ in $\mathbb{R}^n$, they determine a **simplex**

$$C = \text{conv}\left(\{x_i\}_{i=0}^k\right) = \left\{x = \sum_{i=0}^k t_i x_i : t_i \geq 0, \sum_{i=0}^k t_i = 1\right\}.$$

Note. The affine dimension of $k+1$ point simplex is k, so we call it a k-dim simplex.

Intuition. Due to affine independence, we can think of the convex hull of the points as having "all its fat trimmed". Using the rubber band visualization, we can see why the following examples are true.

Example.

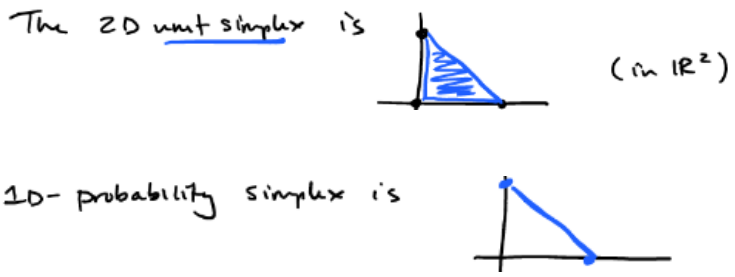
- 1-dim simplex $= \text{conv}(\{x_0, x_1\})$ is a line segment in $\mathbb{R}^n, n \geq 1$.
- 2-dim simplex $= \text{conv}(\{x_0, x_1, x_2\})$ is a filled triangle in $\mathbb{R}^n, n \geq 2$.
- 3-dim simplex is a tetrahedron in $\mathbb{R}^n, n \geq 3$. Clearly if $n = 2$, 4 points cannot be affinely independent and thus cannot generate a simplex.
- In $\mathbb{R}^n$, the **unit simplex** is the simplex generated by $\{0, e_1, e_2, \ldots, e_n\}$, where e_i is the standard basis. It has affine dimension of n. This can be expressed as

$$\left\{x \in \mathbb{R}^n : x \geq 0, \sum_{i=1}^n x_i \leq 1\right\} \text{ or } \{x \in \mathbb{R}^n : x \geq 0, \mathbb{1}^T x \leq 1\}.$$

- In $\mathbb{R}^n$, the $(n-1)$-dim **probability simplex** is generated by $\{e_1, e_2, \ldots, e_n\}$

(basically unit simplex without 0). It can be expressed as

$$\{x \in \mathbb{R}^n : x \geq 0, \mathbb{1}^T x = 1\}.$$



Remark. Simplex is generated by finite points. *Atomic norm* generalizes this to infinite points and uses gauge functions for signal processing.

1.3 Operations that preserve convexity

1) Cartesian products: If $C_1 \subseteq \mathbb{R}^{n_1}$ and $C_2 \subseteq \mathbb{R}^{n_2}$ both convex, then $C_1 \times C_2 \subseteq \mathbb{R}^{n_1+n_2}$ is convex.

2) Arbitrary intersections (even uncountable): If C_1, C_2 are convex, then $C_1 \cap C_2$ is convex. *This is not true for unions.*

3) Image and preimage of an affine function $f(x) = Ax + b$:

   - $f(C)$ is convex if $C \subseteq \mathbb{R}^n$ is.
   - $f^{-1}(C)$ is convex if $C \subseteq \mathbb{R}^m$ is.

   This implies that scaling, translation, rotation, and projection all preserve convexity.

   So does **Minkowski sum**:

$$C_1 + C_2 \coloneqq \{x + y : x \in C_1, y \in C_2\}.$$

   We should think of it like a convolution, where each element in C_1 is convolved with the entire C_2.

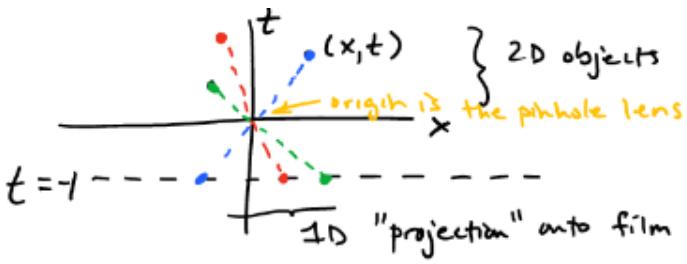
1.3.1 Linear-fractional and perspective functions

Definition: perspective function

A **perspective function** is a function $P : \mathbb{R}^{n+1} = \mathbb{R}^n \times R_{++} \to \mathbb{R}^n$ s.t. for $z \in \mathbb{R}^n$ and $t \in \mathbb{R}_{++}$,

$$P(z,t) = \frac{z}{t}.$$

Intuition. We can think of it as normalizing by t, $(\frac{z}{t}, 1)$, and then projecting to $\mathbb{R}^n$ (or equivalently dropping the last component). Geometrically we can think of it as a "pin-hole" camera that projects 3D points in the t-positive half of $\mathbb{R}^3$ through the pin-hole at origin onto the 2D film at $t = -1$. This gives us $(-\frac{z}{t}, -1)$ which is the negative perspective. Then since all the points are on $t = -1$ we simply drop it.



If P is the perspective function, we can conclude that

- If $C \in \mathbb{R}^{n+1}$ is convex, then $P(C)$ is convex in $\mathbb{R}^n$.
- If $C = \mathbb{R}^n$ is convex, then $P^{-1}(C)$ is convex in $\mathbb{R}^{n+1}$.

Definition: linear-fractional function

A **linear-fractional function** is $f : \mathbb{R}^n \to \mathbb{R}^m$ that composes the perspective P with an affine function g, where

$$g(x) = \begin{pmatrix} A \\ c^T \end{pmatrix} x + \begin{pmatrix} b \\ d \end{pmatrix} : \mathbb{R}^n \to \mathbb{R}^{m+1}, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n, d \in \mathbb{R}.$$

That is, $f = P \circ g$, and

$$f(x) = \frac{Ax + b}{c^T x + d}, \text{ with domain } \{x : c^T x + d > 0\}.$$

Note. Since the image and preimage of both affine and perspective functions preserve convexity, the image and preimage of a linear-fractional function again preserve convexity.

1.3.2 Generalized inequalities

Definition: proper cone

A cone $K \subseteq \mathbb{R}^n$ is called a **proper cone** if it satisfies the following:

1) convex
2) closed
3) solid or nonempty interior
4) pointed or contains no line or $x \in K, -x \in K \Rightarrow x = 0$.

Proposition

Any proper cone K induces a partial order:

$x \preceq_K y$ or simply $x \leq y$ if $y - x \in K$,

$x \prec_K y$ or simply $x < y$ if $y - x \in \text{int}(K)$.

Definition: dual cone

If K is a set, its **dual cone** is

$$K^* = \{y : \langle x, y \rangle \geq 0 \ \forall \ x \in K\}.$$

Note. The larger K is, the more restricted K^* becomes.

Properties of dual cones

1) K^* is a cone even if K isn't a cone.
2) K^* is convex even if K isn't convex.
3) $K_1 \subseteq K_2 \Rightarrow K_2^* \subseteq K_1^*$.
4) $K^{**} = K$ iff K is a proper cone.

Example. If K is a subspace, then $K^* = K^\perp$. This is because $-x \in K$ so equality is achieved in the definition.

Example (self-dual)**.** $\mathbb{R}^n_+ = \left(\mathbb{R}^n_+\right)^*$ because it is a proper cone.

Example (PSD matrices)**.** $K = S^n_+$, and $x \in K \Rightarrow X = GG^T$ (Cholesky).

Then

$$K^* = \{Y \in S^n : \langle Y, X \rangle \geq 0 \ \forall \ X \succeq 0\}.$$

Recall that

$$\begin{aligned}
\langle Y, X \rangle &= \text{tr}(Y^T X) \\
&= \text{tr}(YX) \text{ since } Y = Y^T \\
&= \text{tr}(YGG^T) \\
&= \text{tr}(G^T Y G) \text{ by cyclic property of trace}
\end{aligned}$$

The last expression is ≥ 0 for all matrices G iff $Y \succeq 0$. Hence we show that S^n_+ is self-dual.

1.3.3 separating and supporting hyperplanes

Theorem: separating hyperplane

Let C, D be convex, non-intersecting sets in $\mathbb{R}^n$, then there exists $a \in \mathbb{R}^n \setminus \{0\}$ and $\mu \in \mathbb{R}$ s.t.

$$a^T x \leq \mu \ \forall \ x \in C$$
$$a^T x \geq \mu \ \forall \ x \in D$$

Note. This reads as there exists a hyperplane that separates the two convex sets. It is clearly not true if the sets aren't convex. a is the normal to the hyperplane.

Definition: Chebyshev set

A set S is a **Chebyshev set** if for all x_0, there exists a unique $x \in S$ s.t.

$$x = \underset{y \in S}{\text{argmin}} \, \|y - x_0\|.$$

Note. This reads as there exists a unique best approximation point in the set S for any x_0.

Example. Open unit ball isn't Chebyshev because it doesn't reach infimum.

Example. A nonconvex set isn't Chebyshev because there exists an x_0 where we have at least two best approximation points.

Theorem

Any nonempty, closed, convex set in a Hilbert space is Chebyshev.

Theorem: supporting hyperplanes

(i) If C is convex, closed and $D = \{x_0\}, x_0 \notin C$, then there exists $a \in \mathbb{R}^n$ s.t. $a^T x < a^T x_0 \ \forall \ x \in C$.

(ii) Same but C needs not be closed, $x_0 \notin \overline{C}$.

(iii) as in (ii) but allow $x_0 \in \overline{C} \setminus C$.

Proof: (i)

WLOG let $x_0 = 0$ (since we can always translate C). C is Chebyshev so let y be the unique closest point to 0, and define $a = -y$ (normal of the hyperplane). We wish to show that $a^T x < a^T x_0 = 0 \ \forall \ x \in C$. That is, $y^T x > 0 \ \forall \ x \in C$.

Given $x \in C$, $y + \varepsilon(x - y) \in C$ by convexity. Since y is the best approximation point,

$$\begin{aligned} \|y\|^2 &\leq \|y + \varepsilon(x-y)\|^2 \\ &= \|y\|^2 + 2\varepsilon\langle y, x-y\rangle + \varepsilon^2\|x-y\|^2 \\ 0 &= 2\langle y, x\rangle - 2\langle y, y\rangle + \varepsilon\|x-y\|^2 \\ \langle y, x\rangle &\geq \|y\|^2 - \frac{\varepsilon}{2}\|x-y\|^2 \end{aligned}$$

Take $\varepsilon \to 0$, since $y \neq 0 \Rightarrow \|y\| > 0$, we obtain $y^T x > 0$ as required. □

Remark. This is related to **Theorems of Alternatives**. Generally, they are stated as the following:

Either A is true, B is false, but not both.

Example (Fredhold alternative,finite-dim)**.** Either $\{x : Ax = b\}$ is empty, or $\{\lambda : A^T\lambda = 0, \lambda^T b \neq 0\}$ is non-empty, but not both.

Why do we care? To prove that there is a solution to $Ax = b$. We can simply find a solution x. This is a "certificate". But if professor asks you to prove there isn't a solution to $Ax = b$, we can try to show that A is singular, but if $b = 0$ even singular A works. Another way is to find a "certificate" λ. This is the first task of duality.

Example (Farkas Lemma)**.** Either $\{Ax = b, x \geq 0\}$ is non-empty, or $\{\lambda : A^T\lambda \geq 0, \lambda^T b < 0\}$ is non-empty, but not both.

Theorem: Thereom of Alternatives for strict linear inequalities

The following statements are equivalent:

(i) The set $\{x : Ax < b\}$ is empty.

(ii) The sets $C = \{b - Ax : x \in \mathbb{R}^n\}$ and $D = \mathbb{R}^m_{++}$ do not intersect.

(iii) The hyperplane separation theorem and its converse hold. That is,

$$\exists\ \lambda \geq 0 (\lambda \neq 0) \text{ s.t. } A^T\lambda = 0, \lambda^T b \leq 0.$$

Intuition. (ii) is just rephrasing (i). No intersection from (ii) can then be established by finding something that separates C, D in (iii).

Proof: converse of hyperplane separation

(iii) $\Rightarrow$ (i): suppose such λ exists, and for contradiction, assume there exists x s.t. $Ax < b$. Then since $\lambda \geq 0$,

$$0 = (A^T\lambda)^T x = \lambda^T Ax < \lambda^T b.$$

So we obtain $0 < \lambda^T b \leq 0$, a contradiction.

(i) $\Rightarrow$ (iii): By the separation theorem, we know there exists $\lambda \neq 0$ s.t.

$$\lambda^T(b - Ax) \leq \mu, x \in \mathbb{R}^n$$
$$\lambda^T y \geq \mu, y \in \mathbb{R}^n_{++}$$

It follows from the first condition that $\lambda^T Ax = 0$ because otherwise we can just choose a large negative x to exceed μ and get contradiction. Since this is true for all x, it must be that $\lambda^T A = A^T\lambda = 0$. From the second condition we have $\lambda \geq 0$, because otherwise if $\lambda_i < 0$, we can choose $y_i \to \infty$ to get contradiction. Moreover, we need $\mu \leq 0$ since if $\mu > 0$, we can take all components of y to 0^+, so $\lambda^T y \to 0^+$. Then $\lambda^T(b - A^T x) \leq \mu \leq 0$ implies that $\lambda^T b \leq 0$.

Taken together, we have $\lambda \geq 0, \lambda \neq 0$, $A^T\lambda = 0$, and $\lambda^T b \leq 0$. $\square$

1.4 Convex Functions [BV04 Ch.3]

Definition: convex function

A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex** if $\text{dom}(f)$ is a convex set and for all $x, y \in \text{dom}(f)$ and $0 \le t \le 1$,

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y).$$

It is **strictly convex** if it has strict inequality. It is **strongly convex w.r.t. the norm $\|\cdot\|$ with parameter μ** if $\text{dom}(f)$ is a convex set and, for all $x, y \in \text{dom}(f), x \ne y, 0 \le t \le 1$,

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) - \frac{\mu}{2}t(1-t)\|x-y\|^2.$$

Intuition. For strongly convex, we can see that when $t = \frac{1}{2}$ or when the point is midway between x, y, the bound of inequality is a lot smaller than when t is closer to either point. This forces the function to have a large curvature.

Theorem: simpler characterizations

f is convex if $\text{epi}(f)$ is convex ($\Rightarrow \text{dom}(f)$ is convex).

f is strictly convex means it always has curvature and no straight lines.

f is strongly convex with parameter μ and w.r.t. $\|\cdot\|_2$ iff $x \mapsto f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex (not true for general norms).

Note. Subtraction of convex function doesn't preserve convexity, except in the case of strongly convex w.r.t. Euclidean norm.

Remark. Convexity is a *global property.* This contrasts with continuity which is a local property.

Remark. In convex analysis, we allow the *extended value function* $f : \mathcal{H} \to [-\infty, \infty]$ or $f : \mathcal{H} \to (-\infty, \infty]$, where $\mathcal{H}$ is a generic Hilbert space.

This way, if $x \in \text{dom}(f)$, we can pretend it is but define $f(x) = +\infty$. This

wouldn't affect our minimization problem. Now we can redefine

$$\text{dom}(f) = \{x : f(x) < +\infty\}.$$

This will turn out to be convenient for minimization.

Example. Define the *indicator function* of a set C to be

$$I_C(x) = \begin{cases} 0, & x \in C \\ +\infty, & x \notin C \end{cases}$$

This is different than how we usually define indicator function. Now we can do the following:

$$\min_{x \in C} f(x) \Leftrightarrow \min_{x \in \mathbb{R}^n} f(x) + I_C(x).$$

That is, we can turn a constrained minimization problem into an unconstrained problem with huge penalty on going outside the constraint.

Definition: proper function

$f : \mathbb{R}^n \to [-\infty, \infty]$ is **proper** if

1) it never takes the value $-\infty$.

2) $\text{dom}(f) \neq \emptyset$. That is, the value doesn't always equal to $+\infty$.

Note. This way we can assume there exist feasible points, hence "proper".

Example. I_C is proper iff $C \neq \emptyset$.

Definition: lower semi-continuous (lsc)

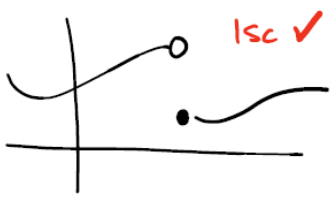
$f : \mathbb{R}^n \to [-\infty, \infty]$ is **lower semi-continuous (lsc)** at $x \in \mathbb{R}^n$ if for all (X_n) s.t. $x_n \to x$,

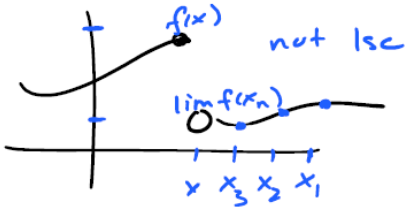
$$f(x) \leq \liminf_n f(x_n) \coloneqq \lim_{n \to \infty} \inf_{n \leq k} f(x_k).$$

Intuition. This is a lot like sequential definition of continuity.

Definition: lsc function

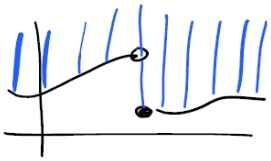
f is a **lsc function** if f is lsc for all points $x \in \mathbb{R}^n$.

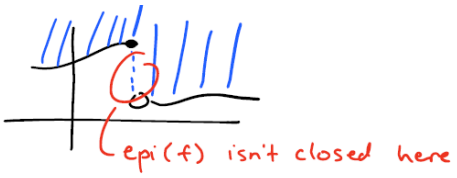




Theorem

In $\mathbb{R}^n$ or any Hausdorff space, f is lsc iff epi(f) is a closed set.





Example. I_C is lsc iff C is a closed set.

We can extend classical theorems involving continuity such as the Extreme Value Theorem to lsc.

Theorem

If C is compact, then f is lsc $\Rightarrow$ f achieves its minimum over C.

Remark. f is continuous iff f is lsc and usc.

Notation. $\Gamma(\mathbb{R}^n)$ is the set of all lsc and convex functions $f : \mathbb{R}^n \to [-\infty, \infty]$.

$\Gamma_0(\mathbb{R}^n) \subseteq \Gamma(\mathbb{R}^n)$ consists of such functions that is also proper, $f : \mathbb{R}^n \to (-\infty, \infty]$.

So $\Gamma_0(\mathbb{R}^n)$ is the standard class of functions for convex optimization.

Example. $I_C \in \Gamma_0(\mathbb{R}^n)$ for some $C \subseteq \mathbb{R}^n$ iff C is nonempty, closed, and convex.

Remark. Recall that the restriction to proper function is mild.

What about restricting to lsc functions? It's also mild in the context of convex functions, because "weird things with convex functions can only involve boundaries (and $+\infty$).

Theorem: 8.38 [BC17

] If $f : \mathcal{H} \to (-\infty, \infty]$ is proper and convex, then f is continuous at $x \in \text{dom}(f)$ iff f is bounded above on a neighborhood of x.

Note. For convex functions, we won't see jumps like in the lsc case.

Corollary: 8.39

Given the same setup,

if f is bounded above on some neighborhood of x, or

if f is lsc, or

if $\mathcal{H}$ is finite dimensional,

then f is continuous on the interior of its domain, $\text{int}(\text{dom}(f))$.

Note. Under these assumptions, weird discontinuous things can only happen at the boundary.

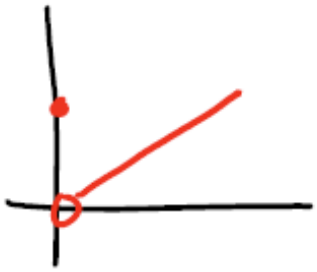


Figure 1.1: An example of a proper, convex function (not lsc) that isn't continuous due to discontinuity at the boundary. It is however continuous on the interior.

Remark. In summary, by the corollary if $f : \mathbb{R}^n \to \mathbb{R}$ (has full-domain and not equal to $\pm\infty$), then convex $\Rightarrow$ continuous.

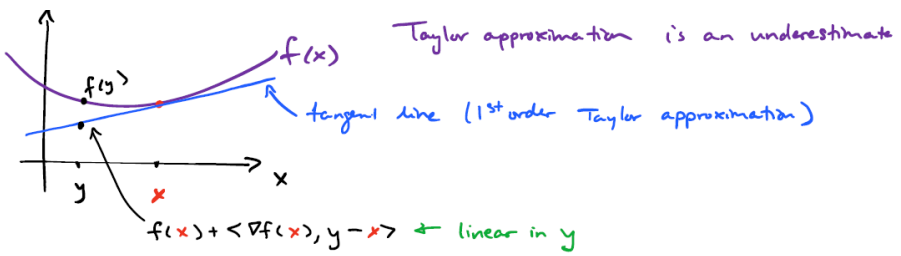
1.4.1 First-order conditions

Theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on dom(f) and if dom(f) is open and convex, then f is convex iff for all $x, y \in \text{dom}(f)$,

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

Note. This is the 1st order Taylor approximation (tangent line). The line is supporting the epigraph of f.



Theorem

Under the same assumption, f is convex iff ∇f is monotone. That is, for all $x, y \in \text{dom}(x)$,

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

Intuition. Recall in 1D, f is convex if slope is non-decreasing. That is, if $x - y \geq 0$, then $f'(x) - f'(y) \geq 0$ and if $x - y \leq 0$ then $f'(x) - f'(y) \leq 0$. A concise way to express that is $(x - y)(f'(x) - f'(y)) \geq 0$. Here we generalize this to higher dimensions.

Theorem: 2nd-order condition

$f : \mathbb{R}^n \to \mathbb{R}$. If the Hessian $\nabla^2 f(x)$ exists for all $x \in \text{dom}(f)$, then

a) f is convex iff $\nabla^2 f(x) \succeq 0 \ \forall \ x \in \text{dom}(f)$.

b) f is μ-strongly convex (w.r.t. $\|\cdot\|_2$) iff $\nabla^2 f(x) \succeq \mu I$.

If $\nabla^2 f(x) \succ 0$, then f is **strictly convex**.

Remark. f can be convex but $\nabla f, \nabla^2 f$ need not exist!

What if f isn't differentiable?

Definition: subdifferential

Let $f : \mathbb{R}^n \to (-\infty, \infty]$ be proper, then we define the **subdifferential** of f at x to be

$$\partial f(x) = \{d \in \mathbb{R}^n : \ \forall \ y \in \mathbb{R}^n, f(y) \geq f(x) + \langle d, y - x \rangle\}.$$

Note. d here is called a **subgradient**.

Theorem

If f is proper and convex then

$$x \in \text{ri}(\text{dom}(f)) \Rightarrow \partial f(x) \neq \emptyset.$$

Note. The proof is related to separating/supporting hyperplanes.

Proposition

$\partial f(x)$ is a singleton iff f is differentiable at x.

Example. $f(x) = |x|$. Then if $x \neq 0$, $f'(x) = \operatorname{sgn}(x)$ and $\partial f(x) = \{f'(x)\}$. If $x = 0$, $f'(0)$ DNE. But $\partial f(0) = [-1, 1]$.

Theorem: Fermat's Rule

If f is a proper function, then

$$\operatorname*{argmin}_x f(x) = \{x : 0 \in \partial f(x)\}.$$

Proof

This just means that we can plug 0 into the definition of subdifferential and get

$$f(y) \geq f(x) + \langle 0, y - x \rangle = f(x) \ \forall \ y.$$

This clearly shows that x is a global minimizer. □

Note. This generalizes the calculus idea of critical points for smooth functions.

Remark. Subdifferentials are a global notion (for all y) whereas gradients are a local notion. How do we reconcile that subdifferential can be the gradient? The answer is that the global property of convexity links the two.

Remark. So all we need to do is to invert ∂f. That is,

$$\operatorname{argmin} f(x) = \partial f^{-1}.$$

In fact, this is usually not practical or even possible especially for interesting problems. It may be possible for subproblems.

Definition: normal cone

The **normal cone** to a set C at point x is

$$N_C(x) = \begin{cases} \{d : \langle d, y - x \rangle \leq 0 \ \forall \ y \in C\} & \text{if } x \in C \\ \emptyset & \text{if } x \notin C \end{cases}$$

Example. Let $C \neq \emptyset$ be convex, so I_C is a proper convex function. Then $\partial I_C = N_C$.

Example. $x \in \operatorname{int} C \Rightarrow N_C(x) = \{0\}$. Why? WLOG, shift C so $x = 0$. If $\langle d, y \rangle \leq 0 \ \forall \ y \in C$. Then $x \in \operatorname{int} C \Rightarrow$ we can choose $y = \varepsilon d \in C$ for sufficiently small $\varepsilon > 0$. Then $\varepsilon \|d\|^2 \leq 0 \Rightarrow d = 0$.

Example. $x \in \partial C$ (the boundary). We want d s.t. $\langle d, y \rangle \leq 0 \ \forall \ y \in C$. Geometrically this means we want the angle between d, y to be perpendicular or obtuse. If the boundary is smooth, since d needs to be at least perpendicular to any y immediately to the left and right of x, it must be the normal ray of the tangent plane.

Example.

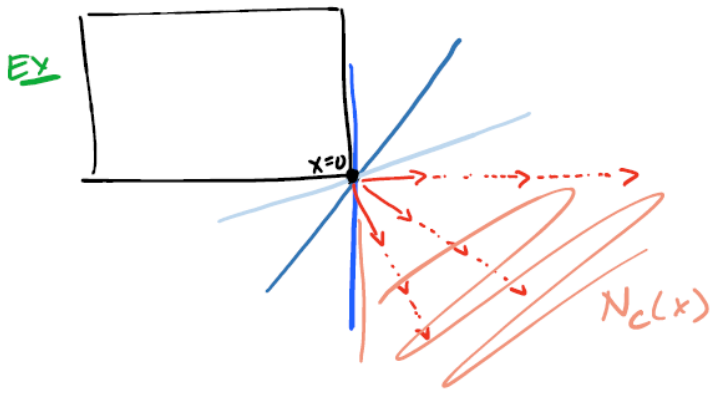


Figure 1.2: The normal cone at non-smooth boundary looks indeed like a cone.

Remark. An equivalent definition of normal cone is the set of all vectors that define a supporting hyperplane to C, passing through x.

Example. If C is a vector space, since C is closed under inverses, if we use $-y$ in addition to y in the definition we will get an equality which implies

orthogonality. Hence

$$N_C(x) = \begin{cases} C^\perp & x \in C \\ \emptyset & x \notin C \end{cases}$$

Proposition: 6.47 BC17

If $C \neq \emptyset$ is closed and convex, then $x = P_C(y)$ iff $y - x \in N_C(x)$, where $P_C(y)$ denotes the orthogonal projection of y onto C.

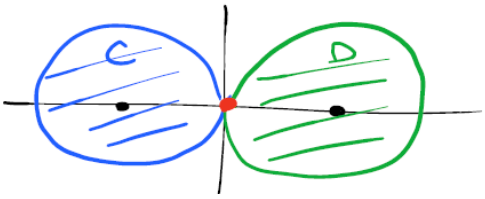
1.4.2 Calculus

Remark. Calculus is a set of rules we can use to calculate.

One such rule is that derivatives/gradients are linear.

Is it true that $\partial(f+g) = \partial f + \partial g$, where $+$ is the Minkowski sum? No! Although it's often true.

Example. $f = I_C, g = I_D \in \mathbb{R}^2$.



C, D balls of radius 1 centred at (−1, 0) and (+1, 0), resp.

Then $\partial(f+g)(x) = \partial f(x) \partial g(x)$ for all x except at $x = 0$. At $x = 0$, recall that

$$\partial f(0) = N_C(0) = \mathbb{R}_+ \times \{0\}$$
$$\partial g(0) = N_D(0) = \mathbb{R}_- \times \{0\}$$

So $\partial f(0) + \partial g(0) = \mathbb{R} \times \{0\}$ But

$$\begin{aligned} \partial(f+g)(0) &= N_{C \cap D}(0) && \text{by def of indicator} \\ &= N_0 \\ &= \{d : \langle d, y - 0 \rangle \leq 0 \ \forall \ y \in \{0\}\} && \text{vacuous constraint} \\ &= \mathbb{R}^2 \end{aligned}$$

We can see this counterexample is somewhat contrived, so linearity is often true.

Remark. Sufficient conditions to guarantee when this linearity is true are called **constraint qualifications (CQ)**.

Corollary: 16.48 (iv) BC17

If $f, g \in \Gamma_0(\mathcal{H})$, and $\mathcal{H} = \mathbb{R}^n$, and one of the following holds:

(i) $\operatorname{ri}(\operatorname{dom}(f)) \cap \operatorname{ri}(\operatorname{dom}(g)) \neq \emptyset$.

(ii)

$dom(f) \cap \text{int}(\text{dom}(g)) \neq \emptyset$.

(iii) either f or g has full domain (all of $\mathbb{R}^n$).

Note. (iii) is most commonly used.

Since the previous example didn't satisfy a CQ, the linearity didn't hold. That is, $\text{dom}\, f = C, \text{dom}\, g = D, \text{int}\, C \cap \text{int}\, D = \emptyset$.

Remark. There are other cones including **tangent, polar, recession/asymptotic, and barrier cones**.

1.4.3 Lipschitz gradient

An easier way to show F is Lipschitz-continuous: if F' exists, then $|||F'||| \leq L \Rightarrow F$ is Lipschitz continuous (by the definition of derivative/Jacobian and some manipulation).

Notation. $|||\cdot|||$ denotes the appropriate operator norm, usually spectral norm if the original norm is Euclidean.

Remark. In optimization, "Jacobian" is often confusing, since it's unclear what F is. Of the objective function or of the gradient? Instead we prefer to say the Jacobian of the objective is the gradient (transposed). The Jacobian of the gradient is the Hessian.

Remark. The Hessian can be thought of as a bilinear operator $\langle d, \nabla^2 f(x) d \rangle$

Theorem

Suppose convex $f \in \mathcal{C}^2(U)$ for some open set $U \subseteq \mathbb{R}^n$, then

$$\nabla f \text{ is } L\text{-Lipschitz continuous on } U \Leftrightarrow \ \forall\, x \in U, \nabla^2 f(x) \preceq LI.$$

That is, all eigenvalues of $\nabla^2 f(x) \leq L \Rightarrow |\|\nabla^2 f(x)\|| \leq L$.

Theorem

Same setup, then

f is μ-strongly convex on $U \Leftrightarrow \ \forall\, x \in U, \mu I \preceq \nabla^2 f(x)$.

Note. We assume $\mu > 0$ since $\mu = 0$ would give us plain old convexity.

Remark. One of our common assumption will be ∇f is L-Lipschitz continuous ($\nabla^2 f \preceq LI$) and a bit less common, also assume strong convexity ($uI \preceq \nabla^2 f$).

Example (best function ever)**.** Consider $f(x) = \frac{1}{2}\|x\|_2^2, \nabla f(x) = x, \nabla^2 f(x) = I$. So $L = 1, \mu = 1$. This is the only function with this property.

This is the nicest function ever for optimization!

Definition: condition number

The **condition number** of f is $k_f = \frac{L}{\mu}$. $k_f \approx 1$ is good. Larger is bad.

Why do we care about these assumptions?

Recall from calculus, Taylor's theorem states that

$$f(y) = f(x) + f'(x)(y-x) + \frac{1}{2}f''(\xi)(y-x)^2,$$

were $\xi \in [x, y]$. If $f''(\xi) \leq L \; \forall \; \xi$, then

$$f(y) \leq f(x) + f'(x)(y-x) + \frac{L}{2}(y-x)^2.$$

Theorem

If ∇f is L-Lipschitz continuous and f is μ-strongly convex, then for all $x, y \in \text{dom}(f)$,

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

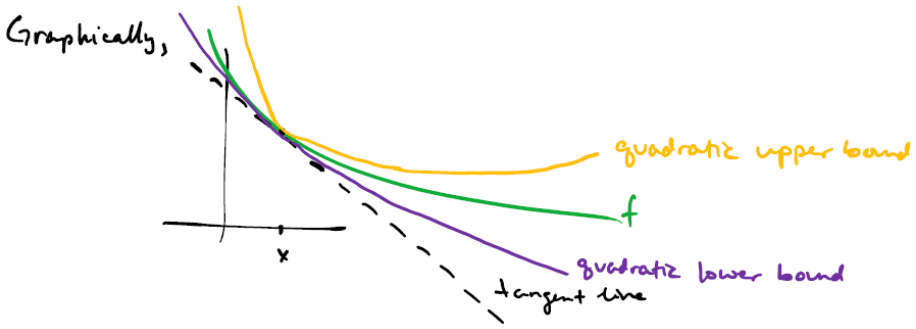


Figure 1.3: If f is complicated but we can "sandwich" it between a quadratic upper bound and a quadratic lower bound ($\mu > 0$) or a linear lower bound ($\mu = 0$), then we can work with the quadratics to understand the behavior of f since quadratics are much easier to deal with.

See more properties from this section in the Github handout StrongConvex-

ityLipschitz.pdf.

1.4.4 Examples [BV04 Ch.3.1.5]

Examples of convex functions $f : \mathbb{R} \to \mathbb{R}$:

- $e^{ax}, a \in \mathbb{R}$.
- x^a on $x \in \mathbb{R}_{++}$ if $a \leq 0$ or $a \geq 1$. (It's concave on $0 \leq a \leq 1$).
- $|x|^a$ on all of $\mathbb{R}$, if $a \geq 1$.
- $-\log_b(x)$ on $\mathbb{R}_{++}$ if $b > 1$.
- On $\mathbb{R}^+$,
$$\begin{cases} x \cdot \log(x) & x > 0 \\ 0 & x = 0 \end{cases}$$
since $f''(x) = \frac{1}{x} > 0$.

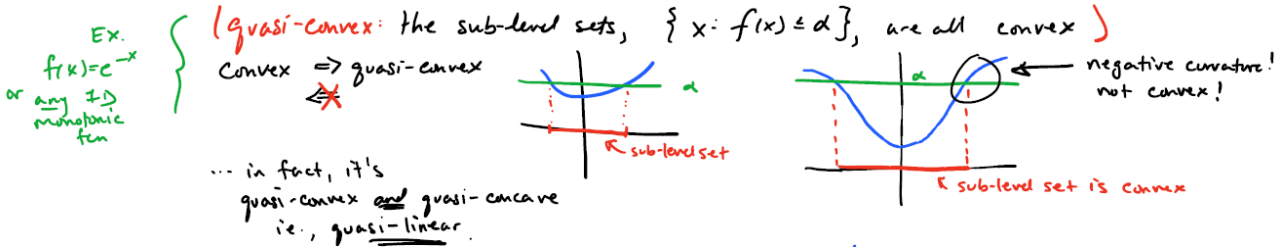
Examples of convex functions $f : \mathbb{R}^n \to \mathbb{R}$:

- any norm/seminorm (follows directly from triangle inequality).
- $f(x) = \max\{x_1, \ldots, x_n\}$.
- $f(x, y) = x^2/y$, $\mathrm{dom}(f) = \mathbb{R} \times \mathbb{R}_{++}$. "Quadratic over linear".

  $f(x, y) = \|x\|_2^2/y$, $\mathrm{dom}(f) = \mathbb{R}^{n-1} \times \mathbb{R}_{++}$.

  $f(x, Y) = x^T Y^{-1} x$, $\mathrm{dom}(f) = \mathbb{R}^n \times S^n_{++}$. "Matrix fractional function".

  Note. "Linear fractional function"
$$g(x) = \frac{Ax + b}{c^T x + d}, \quad \mathrm{dom}(g) = \{x : c^T x + d > 0\}$$
  is not convex but it is **quasi-convex**. It is defined by having all convex sub-level sets $\{x : f(x) \leq \alpha\}$.



- "log-sum exp" aka "soft-max"

$$f(x) = \frac{1}{\alpha} \log\left(e^{\alpha x_1} + \ldots + e^{\alpha x_n}\right), \alpha > 0.$$

  This is differentiable but needs to be careful about numerical under/overflow.

- geometric mean $f(x) \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$ on $\mathbb{R}^n_{++}$.
- $-\log\det(X) = -\log\left(\prod \lambda_i\right) = -\sum \log(\lambda_i)$ on S^n_{++}.

Theorem: Jensen's Inequality

$$f(\mathrm{E}[x]) \leq \mathrm{E}[f(x)].$$

Remark. Let X be a random variable that outputs points in $\mathrm{dom}(f)$ with probability in $[0, 1]$, then the inequality follows from definition of convex function.

Example. In machine learning, we often prove something like

$$\mathrm{E}[\| \text{ error}\|^2] \leq \varepsilon.$$

Let $f(x) = x^2$. So by Jensen's inequality:

$$(\mathrm{E}[\| \text{ error}\|])^2 \leq \mathrm{E}\left[\| \text{ error}\|^2\right] \leq \varepsilon$$
$$\mathrm{E}[\| \text{ error}\|] \leq \sqrt{\mathrm{E}\left[\| \text{ error}\|^2\right]} \leq \sqrt{\varepsilon}$$

Recall that $\| \text{ error}\|^2$ is the nicest function ever.

Remark. Hölder's inequality/Cauchy-Schwarz can also be proved via Jensen.

Theorem: Hölder's inequality

If $\frac{1}{p} + \frac{1}{q} = 1$,

$$|\langle x, y\rangle| \leq \|x\|_p \cdot \|y\|_q.$$

Remark. We can use Jensen's to prove Holder inequality.

1.4.5 Preserving convexity

Rule 0: non-negative (weighted) sums

If $f_1, \ldots, f_m$ are convex, $\alpha_i \geq 0$, then $x \mapsto \sum \alpha_i f_i(x)$ is convex too.

Subtraction (negative weights) doesn't work.

It works for integrals too:

If for all y, $f(\cdot, y)$ is convex, and $w(y) \geq 0$. Then

$$x \mapsto \int_\Omega f(x,y)w(y)\, dy$$

is convex.

Rule 1: perspective function

Definition: perspective

Let $f : \mathbb{R}^n \to \mathbb{R}$, then its **perspective** is $g : \mathbb{R}^{n+1} \to \mathbb{R}$,

$$g(x,t) = t \cdot f\left(\frac{x}{t}\right), \quad \text{dom}(g) = \{(x,t) : x/t \in \text{dom}(f), t > 0\}.$$

Proposition

$f : \mathbb{R}^n \to \mathbb{R}$ is convex $\Rightarrow$ its perspective is convex.

Example. $f(x) = \|x\|^2$ is convex. Its perspective is

$$t \cdot \left\|\frac{x}{t}\right\|^2 = t \cdot \frac{\|x\|^2}{t^2} = \frac{\|x\|^2}{t}.$$

This is the quadratic-over-linear example we saw earlier. This is the proof that it is convex.

Example. $f(x) = -\log(x)$ is convex. Its perspective is

$$-t \cdot \log\left(\frac{x}{t}\right) = t \cdot \log(t) - t \cdot \log(x), x, t > 0.$$

This is the **relative entropy** of t, x. More generally, the **Kullback-Leibler**

divergence is

$$D_{KL}(u,v) = \sum_{i=1}^{n} u_i \log\left(\frac{u_i}{v_i}\right) - u_i + v_i.$$

This is an example of **Bregman Divergence**, which we often use to measure "distance" as an alternative to metric. It's especially good for probability distributions.

Rule 2: special types of compositions

Composition of convex functions typically doesn't preserve convexity!

Theorem

f is convex if

(i) h is convex and

(ii) if $k = 1$, h is nondecreasing and g is convex or h is nonincreasing and g is concave.

(iii) if $k > 1$, we enforce (ii) to each argument of h and each g_i.

Note. For nonincreasing/decreasing, we must take into account $\pm\infty$, since in convex analysis we assign infinity to any point not in the domain. So although $h(x) = x$ is nondecreasing on $\mathbb{R}$, if we restrict $\text{dom}(h) = [0,1]$ then it is not nondecreasing anymore.

Theorem: tattoo-worthy

$f = h \circ g$ is convex if h is convex and g is affine.

Example. $f(x) = \|Ax - b\|^2$ is convex by this theorem.

Rule 3: min/max

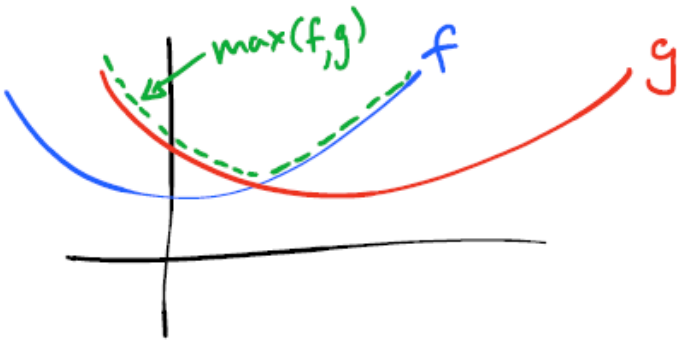
Proposition

If f, g both convex, then $x \mapsto \max\{f(x), g(x)\}$ is convex.

Proof

The epigraph of the maximum is the intersection of two convex epigraphs.

Convex sets are closed under arbitrary intersections. □



Note. This works for supremum too due to closure under arbitrary intersections.

Example.

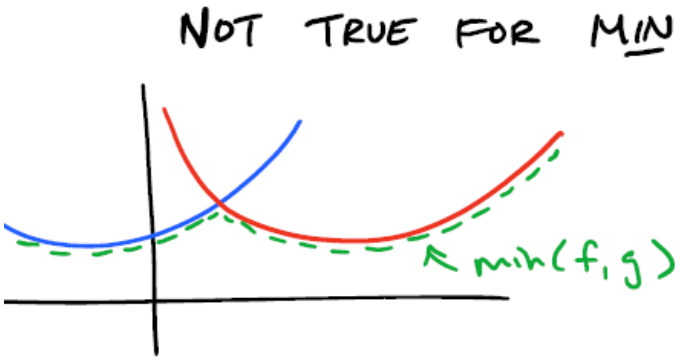
$$f(x) = \sup_{y \in \mathcal{A}} f(x; y)$$

is convex as long as $f(\cdot; y)$ is convex $\forall\ y \in \mathcal{A}$, where $\mathcal{A}$ is an arbitrary set that can be uncountable.

Example (spectral norm)**.**

$$f(A) = \|A\|_\infty = \sup_{\|x\|_2 = 1} \|Ax\|_2$$

is convex since $\forall\ x, A \mapsto \|Ax\|_2$ is convex (composition of convex and affine).



It's easy to see that min doesn't necessarily preserve convexity because it unions epigraphs instead.We need to impose more restrictions to make it work:

Theorem

If $f : \mathbb{R}^n \times \mathbb{R}^m$ is (jointly) convex and if $C \neq \emptyset$ is a convex set, then

$$g(x) = \inf_{y \in C} f(x, y) \text{ is convex.}$$

Example. $\min\{f_1(x), f_2(x)\}$ is not usually convex since this is like taking

$$f(x, y) = \begin{cases} f_1(x), & y = 1 \\ f_2(x), & y = 2 \end{cases}$$

and constraint $C = \{1, 2\}$ is not convex.

Example. The distance to a convex set is a convex function. Let $C \neq \emptyset$ be convex,

$$f(x) = \inf_{y \in C} \|x - y\|.$$

Prove $(x, y) \mapsto \|x - y\|$ is convex.

Proof

We know $z \mapsto \|z\|$ is convex. Consider the linear operator $A(x, y) = x - y$. That is,

$$A \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} I & -I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = x - y.$$

Then the composition of convex and affine is still convex. □

1.4.6 Conjugate functions

Definition: Fenchel-Legendre conjugate

The **F.L. conjugate** of f is

$$f^*(y) = \sup_x \{\langle y, x\rangle - f(x)\}.$$

Note. For matrix inputs, use $\operatorname{tr}\left(Y^T X\right)$.

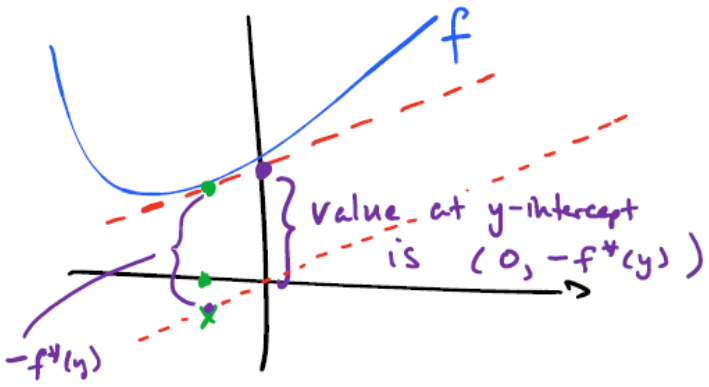
Proposition

f^* is always convex (whether f is or not).

Proof

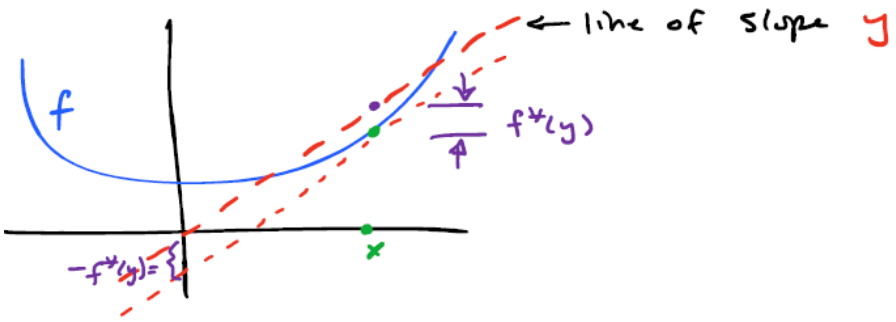
$y \mapsto \langle y, x\rangle - f(x)$ is an affine function of y which is convex, and supremum preserves convexity. □

Example (1D Legendre Transform)**.** Assume f is strictly convex (so f' is strictly monotone/invertible). So $f^*(y)$ is maximized (unique global) when $0 = y - f'(x) \Rightarrow f'(x) = y \Rightarrow x = (f')^{-1}(y)$. We can interpret y as the slope of $f(x)$.

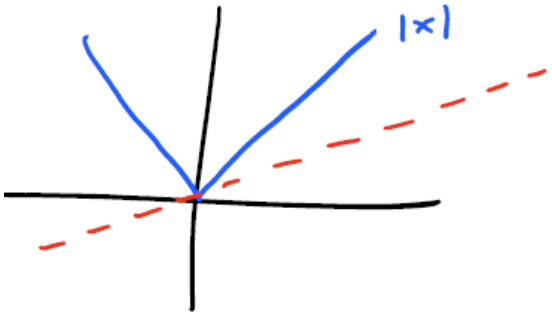


Alternatively, we can think of $f^*(y)$ as finding the x that maximize the signed

separation of the "line: $\langle y, x\rangle$ and f, where the line is on top of f. Then $f^*(y)$ would be the maximized distance.



Example. $f(x) = |x|$. What is $f^*\left(\frac{1}{2}\right)$?



$f^*\left(\frac{1}{2}\right) = 0$ because the line is always below f except at 0.