Convex Optimization and Analysis

Notes by Thomas Gao; transcribed from Professor Moursi's lectures.

Uses the same theorem indexing as the textbook by R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

- Convex Optimization and Analysis
  - Affine Sets and Convex Sets in $\mathbb{R}^n$
    - Best Approximation for Convex Sets
  - Separation Theorems
  - Tangent and Normal Cones
  - More on Convex
    - Subgradient Calculus
    - Calculus of Subdifferentials
    - Differentiability of Convex Function
    - Subdifferentiability and Conjugacy
    - Differentiability and Strong Convexity
  - The Proximal Operator
    - More on Proximal Operator
  - Nonexpansive, Firmly Nonexpansive, and Averaged Operators
  - Constrained Convex Optimization
    - The KKT Conditions
  - Subgradient Method
    - Gradient Descent Classical Theory
    - Projected Subgradient Method
    - The Convex Feasibility Problem
    - Proximal Gradient Method (PGM)
    - The Prox-Grad Inequality
    - Fast Iterative Shrinkage Thresholding Algorithm (FISTA)
    - Iterative Shrinkage Thresholding Algorithm (ISTA)
    - Douglas-Rachford (DR) Operator
    - Stochastic Projected Gradient Method
    - Duality: The Fenchel Duality
    - The Fenchel-Rockafellar Duality
    - DR as a Self-Duel Method

Affine Sets and Convex Sets in $\mathbb{R}^n$

Affine Set: Let $S \subseteq \mathbb{R}^n$. Then

1. S is an **affine set** if $\forall x, y \in S$ and $\lambda \in \mathbb{R}$,

$$\lambda x + (1-\lambda)y \in S$$

   i.e. an affine set contains all lines passing through any two points in the set. Note that trivially, $\emptyset, \mathbb{R}^n$ are affine sets.
2. S is an **affine subspace** if $S \neq \emptyset$ and S is an affine set.
3. Let $S \subseteq \mathbb{R}^n$. The **affine hull** of S, denoted by $\mathrm{aff}(S)$, is the intersection of all affine sets containing S, i.e. the smallest affine set containing S.

Example of affine sets of $\mathbb{R}^n$:

1. L where $L \subseteq \mathbb{R}^n$ is a linear subspace.
2. $a + L$ where $a \in \mathbb{R}^n, L \subseteq \mathbb{R}^n$ is a linear subspace
3. $\emptyset, \mathbb{R}^n$

Convex Set: A subset C of $\mathbb{R}^n$ is convex if $\forall x, y \in C$, and $\forall \lambda \in (0,1)$, we have

$$\lambda x + (1-\lambda)y \in C, \qquad \text{convex combination}$$

The following are examples of convex subsets of $\mathbb{R}^n$:

1. $\emptyset, \mathbb{R}^n$
2. C where C is a ball
3. C where C is an affine set
4. C, where C is a half-space, i.e. $C = \{x \in \mathbb{R}^n : \langle x, u \rangle \leq n\}$ for some fixed $u \in \mathbb{R}^n$.

> Proving Convexity: show that the convex combination of two points in the set must also be in the set.

Theorem 2.1: The intersection of an arbitrary collection of convex sets is convex.

Corollary 2.1.1: Let $b_i \in \mathbb{R}^n, \beta_i \in \mathbb{R}$ for $i \in I$ where I is an arbitrary index set. Then the set

$$C = \{x \in \mathbb{R}^n : \langle x, b_i \rangle \leq \beta_i \forall i \in I\}$$

Proof Sketch: show that half space is convex, and apply Theorem 2.1. □

Convex Combination: A vector sum $\lambda_1 x_1 + \cdots + \lambda_m x_m$ is called a convex combination of vectors $x_1, \ldots, x_m$ if $\forall i \in \{1, \ldots, m\}$, $\lambda_i \geq 0$ and $\sum_{i=1}^{m} \lambda_i = 1$.

Theorem 2.2: a subset C of $\mathbb{R}^n$ is convex if and only if it contains all the convex combination of its elements.

Proof $(\Rightarrow)$: simple induction. □

Convex Hull: Let $S \subseteq \mathbb{R}^n$, the intersection of all convex sets containing S is called the convex hull of S and is denoted by $\mathrm{conv}(S)$. By Theorem 2.1, $\mathrm{conv}(S)$ is convex and it is the smallest convex set containing S.

Theorem 2.3, Convex Combination Theorem: Let $S \subseteq \mathbb{R}^n$. Then $\mathrm{conv}(S)$ consists of all the convex combinations of the elements of S, i.e.

$$\mathrm{conv}(S) = \left\{ \sum_{i \in I} \lambda_i x_i : I \text{ is finite index set}, \right.$$
$$\left. (\forall i \in I)\ x_i \in S, \lambda_i \geq 0, \sum_{i \in I} \lambda_i = 1 \right\}$$

Proof: let $D = \left\{ \sum_{i \in I} \lambda_i x_i : I \text{ is finite index set}, (\forall i \in I)\ x_i \in S, \lambda_i \geq 0, \sum_{i \in I} \lambda_i = 1 \right\}$

We have

- Trivially, $S \subseteq D$
- D is convex (brute force use definition)

- $\text{conv}(S) \subseteq D$ (since D is convex set $\supseteq S$)
- $D \subseteq \text{conv}(S)$ by Theorem 2.2

Best Approximation for Convex Sets

Distance Function: Let $S \subseteq \mathbb{R}^n$. The distance to S is the function

$$\begin{aligned} d_S : \mathbb{R}^n &\longrightarrow [0, +\infty) \\ x &\longmapsto \inf_{s \in S} \|x - s\| \end{aligned}$$

Projection onto a Set: Let $\emptyset \neq C \subseteq \mathbb{R}^n$, let $x \in \mathbb{R}^n$, and let $p \in C$. Then p is a projection of x onto C, if $d_C(x) = \|x - p\|$.

If every point in $\mathbb{R}^n$ has exactly one projection onto C, the projection operator onto C denoted by P_C, is the operator that maps every point in $\mathbb{R}^n$ to its unique projection in C.

> Recall these following definitions:
>
> **Cauchy Sequence**: Let $(x_n)_{n \in \mathbb{N}}$ be a sequence in $\mathbb{R}^n$. Then it is a Cauchy sequence if $\|x_m - x_n\| \longrightarrow 0$ as $\min\{m, n\} \longrightarrow +\infty$. Note that in $\mathbb{R}^n$ every Cauchy sequence converges.
>
> **Function Continuity**: Let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ and let $\bar{x} \in \mathbb{R}^n$. Then f is continuous at $\bar{x}$ if and only if for every sequences $(x_n)_{n \in \mathbb{N}}$ such that $x_n \longrightarrow \bar{x}$, we have $f(x_n) \longrightarrow f(\bar{x})$.
>
> **Theorem L2-a**: let $y \in \mathbb{R}^n$ and let $\|.\|$ be the Euclidean norm on $\mathbb{R}^n$. Then the function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ with $x \longmapsto \|x - y\|$ is continuous.

Auxiliary Lemma L2-1: Let x, y, z be vectors in $\mathbb{R}^n$. Then

$$\|x - y\|^2 = 2\|z - x\|^2 + 2\|z - y\|^2 - 4\left\|z - \frac{x + y}{2}\right\|^2$$

Auxiliary Lemma L2-2: Let $x \in \mathbb{R}^n, y \in \mathbb{R}^n$. Then $\langle x, y \rangle \leq 0$ if and only if $\forall \lambda \in [0, 1], \|x\| \leq \|x - \lambda y\|$

Proof ($\Leftarrow$): Suppose that for every $\lambda \in (0, 1]$ we have $\|x - \lambda y\| \geq \|x\|$. This implies $\langle x, y \rangle \leq (\lambda/2)\|y\|^2$. Taking $\lambda \to 0$ gives the desired result. □

Projection Theorem: Let C be nonempty, closed, convex subset of $\mathbb{R}^n$. Then the following hold:

1. $\forall x \in \mathbb{R}^n$, the projection of x onto C exists and is unique.
2. For every $x \in \mathbb{R}^n$ and every $p \in \mathbb{R}^n$: $p = P_c(x)$ if and only if $p \in C$ and $(\forall y \in C)\langle y - p, x - p \rangle \leq 0$

Proof: For part 1:

- Existence: show that $\exists p \in C$ such that $\|x - p\| = d_C(x)$. Recall that $(\forall x \in \mathbb{R}^n)$

  $$d_C(x) = \inf_{c \in C} \|x - c\|$$

  Therefore, there exists a sequence $(c_n)_{n \in \mathbb{N}}$ in C such that

  $$d_C(x) = \lim_{n \to \infty} \|c_n - x\|$$

  Let $m, n \in \mathbb{N}$. Apply the auxiliary lemma L2-1:

  $$\|c_n - c_m\|^2 = 2\|c_n - x\|^2 + 2\|c_m - x\|^2 - 4\left\|x - \frac{c_n + c_m}{2}\right\|^2$$

  Therefore, as $m \to +\infty, n \to +\infty$,

  $$\|c_n - c_m\|^2 \to 2d_C^2(x) + 2d_C^2(x) - 4d_C^2(x) = 0$$

  Hence $(c_n)_{n \in \mathbb{N}}$ is a Cauchy sequence in C and converges to a point, say $p \in C$ (by the closedness of C). Finally, since Euclidean distance is continuous, we have $d_C(x) = \|x - p\|$.
- Uniqueness: Suppose that $\exists p, q \in C$ that satisfies $\|p - x\| = d_C(x) = \|q - x\|$. Apply auxiliary lemma L2-1, again:

$$0 \leq \|p-q\|^2 = 2\|p-x\|^2 + 2\|q-x\|^2 - 4\underbrace{\left\|x - \frac{p+q}{2}\right\|^2}_{\leq d_C^2(x)}$$
$$\leq 2d_C^2(x) + 2d_C^2(x) - 4d_C^2(x) = 0$$

Hence $\|p-q\| = 0$ and equivalently $p = q$.

For part 2

- $(\Rightarrow)$: We have $p = P_C(x) \iff (\mathrm{I.}) \quad p \in C$ and $(\mathrm{II.}) \quad \|x-p\| = d_C(x)$. Now by definition + some tweak, $(\mathrm{II.}) \iff (\forall y \in C)(\forall \alpha \in [0,1])$

$$\|x-p\| \leq \|x - (\alpha y + (1-\alpha)p)\|$$
$$= \|x - p - \alpha(y-p)\| \qquad (\mathrm{III.})$$

  which by auxiliary lemma L2-2, $(\mathrm{III.}) \iff (\forall y \in C)\langle x-p, y-p\rangle \leq 0$.
- $(\Leftarrow)$: basically the reverse of the forward direction. □

Notes regarding projecting theorem:

- In the absense of closedness. $(\forall x \in \mathbb{R}^n \backslash C)$ the projection of x onto C does not exist.
- In the absence of convexity. For example, on the real line $\mathbb{R}$, consider $C = [-2,-1] \cup [1,2]$. Then both 1,-1 are the projections of 0 onto C.

Example using Projection Theorem: Let $\epsilon > 0$ and let $C = \mathrm{ball}(0,\epsilon) = \{x \in \mathbb{R}^n : \|x\| \leq \epsilon\}$. Show that $(\forall x \in \mathbb{R}^n) P_C(x) = \dfrac{\epsilon}{\max\{\|x\|, \epsilon\}} x$.

Solution: we omit $p \in C$ and show that $(\forall y \in C)\langle x-p, y-p\rangle \leq 0$.

- Case 1: $\|x\| \leq \epsilon \Longrightarrow p = x$ and $\langle x-p, y-p\rangle = 0$.
- Case 2: $\|x\| > \epsilon \Longrightarrow p = (\epsilon/\|x\|)x$, moreover:

$$\begin{aligned}\langle x-p, y-p\rangle &= \langle x - (\epsilon/\|x\|)x, y - (\epsilon/\|x\|)x\rangle \\ &= (1-\epsilon/\|x\|)(\langle x, y\rangle - \epsilon\|x\|) \\ &\leq (1-\epsilon/\|x\|)(\|x\|\|y\| - \epsilon\|x\|) \qquad \text{Cauchy-Schwarz} \\ &\leq (1-\epsilon/\|x\|)(\|x\|\epsilon - \epsilon\|x\|) = 0\end{aligned}$$

And we are done. □

Minkowski Sum of Sets: Let C, D be two subsets of $\mathbb{R}^n$. The Minkowski sum of C and D, denoted by $C + D$ is

$$C + D \coloneqq \{c + d : c \in C, d \in D\}$$

Theorem 3.1, Minkowski sum of Convex Sets: Let C_1, C_2 be convex subsets of $\mathbb{R}^n$. Then $C_1 + C_2$ is convex.

Proposition L3-a: Let C, D be nonempty, closed convex subsets of $\mathbb{R}^n$ such that D is bounded. Then $C + D$ is nonempty, closed and convex.

Proof: We have

- $C \neq \emptyset, D \neq \emptyset \Longrightarrow C + D \neq \emptyset$
- C is convex, D is convex $\Longrightarrow C + D$ is convex by Theorem 3.1.

It remains to show that $C + D$ is closed. To this end, take a convergent sequence $(x_n + y_n)_{n\in\mathbb{N}}$ in $C + D$ such that $(x_n)_{n\in\mathbb{N}}$ lies in C and $(y_n)_{n\in\mathbb{N}}$ lies in D, and $x_n + y_n \to z$. Our goal is to show that $z \in C + D$.

Since D is bounded, using Bolzano-Weierstrass Theorem, we know that there exists a convergent subsequence (y_{k_n}) $y_{k_n} \to y \in D$. Consequently, $x_{k_n} \to z - y \in C$.

That is, $z \in C + \{y\} \subseteq C + D$. □

Counter Example to Theorem 3.1: Let $C_1 = \mathbb{R} \times \{0\}, C_2 = \{(x, y) \in \mathbb{R}^2_{++} : xy \geq 1\}$. Then C_1, C_2 are closed and convex. However, $C_1 + C_2 = \mathbb{R} \times \mathbb{R}_+$ which is convex but **open**.

Reason:

- $(C_1 + C_2 \subseteq \mathbb{R} \times \mathbb{R}_{++})$: indeed, let $(z_1, z_2) \in C_1 + C_2$, then there exists $(y_1, 0) \in C_1$ and $(x_1, x_2) \in C_2$ such that $z_1 = x_1 + y_1$ and $z_2 = x_2$. Note that $z_1 = x_1 + y_1 \in \mathbb{R}$ and $z_2 = x_2 > 0$.
- $(C_1 + C_2 \supseteq \mathbb{R} \times \mathbb{R}_{++})$: Let $(x, y) \in \mathbb{R} \times \mathbb{R}_{++}$. Set $c_1 = \left(x - \frac{1}{y}, 0\right)$, and $c_2 = \left(\frac{1}{y}, y\right)$. Then $c_1 \in C_1, c_2 \in C_2$, and $(x, y) = c_1 + c_2 \in C_1 + C_2$. □

Theorem 3.2: Let C be a convex set, let $\lambda_1, \lambda_2 \geq 0$. Then

$$(\lambda_1 + \lambda_2)C = \lambda_1 C + \lambda_2 C$$

Proof:

- $(\subseteq)$: always true even in the absence of convexity. Let $x \in (\lambda_1 + \lambda_2)C$. Then $\exists c \in C$ such that $x = (\lambda_1 + \lambda_2)c = \lambda_1 c + \lambda_2 c \in \lambda_1 C + \lambda_2 C$
- $(\supseteq)$: we assume that $\lambda_1 + \lambda_2 > 0$ (otherwise the result is trivial). Now by convexity we have:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} C + \frac{\lambda_2}{\lambda_1 + \lambda_2} C \subseteq C$$

Equivalently, $\lambda_1 C + \lambda_2 C \subseteq (\lambda_1 + \lambda_2)C$. □

Closed ball: We write $B(x, \epsilon) = \{y \in \mathbb{R}^n : \|y - x\| \leq \epsilon\}$ We also use the convention $B \coloneqq B(0, 1)$, i.e. the **closed unit ball**.

Let $C \subseteq \mathbb{R}^n$:

- **Interior** of C: $\operatorname{int}(C) = \{x : \exists \epsilon > 0 \text{ such that } x + \epsilon B \subseteq C\}$
- **Closure** of C: $\overline{C} = \bigcap \{c + \epsilon B : \epsilon > 0\}$
- **Relative Interior** of a convex set C is its interior within the affine hull of C:

$$\operatorname{ri}(C) = \{x \in \operatorname{aff}(C) : \exists \epsilon > 0 \text{ such that } (x + \epsilon B) \cap \operatorname{aff}(C) \subseteq C\}$$

Example on the real line:

- $C_1 = \{0\} \subseteq \mathbb{R}$: $\begin{cases} \operatorname{int}(C_1) = \emptyset \\ \overline{C_1} = \{0\} \\ \operatorname{ri}(C_1) = \{0\} \end{cases}$
- $C_2 = [a, b)$: $\begin{cases} \operatorname{int}(C_2) = (a, b) \\ \overline{C_2} = [a, b] \\ \operatorname{ri}(C_2) = (a, b) \end{cases}$

Example in $\mathbb{R}^2$:

- $C_1 = \{(0, 0)\}$: $\begin{cases} \operatorname{int}(C_1) = \emptyset \\ \overline{C_1} = \{(0, 0)\} \\ \operatorname{ri}(C_1) = \{(0, 0)\} \end{cases}$
- $C_2 = [a, b] \times \{0\}$: $\begin{cases} \operatorname{int}(C_2) = \emptyset \\ \overline{C_2} = C_2 \\ \operatorname{ri}(C_2) = (a, b) \times \{0\} \end{cases}$
- $C_3 = [-1, 1] \times [-1, 1]$: $\begin{cases} \operatorname{int}(C_3) = (-1, 1) \times (-1, 1) \\ \overline{C_3} = C_3 \\ \operatorname{ri}(C_3) = \operatorname{int}(C_3) \end{cases}$

Proposition L3-b: Let $C \subseteq \mathbb{R}^n$. Suppose that $\operatorname{int}(C) \neq \emptyset$. Then $\operatorname{int}(C) = \operatorname{ri}(C)$.

Proof: Let $x \in \operatorname{int}(C)$. Then $\exists \epsilon > 0$ such that $B(x, \epsilon) \subseteq C$. Hence, $\operatorname{aff}(B(x, \epsilon)) \subseteq \operatorname{aff}(C)$ (exercise). Since the affine hull of a ball is $\mathbb{R}^n$, $\operatorname{aff}(C) = \mathbb{R}^n$. So the definitions of rel. int. and int. are equivalent. □

Dimension of Convex Set: Let $C \neq \emptyset$ be convex. The dimension of C, denoted $\dim(C)$, is the dimension of the affine hull of C, $\operatorname{aff}(C)$. Suppose

$$L \coloneqq \operatorname{aff}(C) - \operatorname{aff}(C)$$

L is a linear subspace. Indeed, the affine subspace might not pass through the origin. Therefore,

$$\dim(\operatorname{aff}(C)) = \dim L$$

Proposition L3-c: Let C be a convex set in $\mathbb{R}^n$. Then $(\forall x \in \operatorname{int}(C))(\forall y \in \overline{C})$, we have

$$[x, y) \subseteq \text{int}(C)$$

i.e. $(1-\lambda)x + \lambda y \in \text{int}(C),\ \forall \lambda \in [0, 1)$.

Proof: Let $\lambda \in [0, 1)$. We need to show that $(1+\lambda)x + \lambda y + \epsilon B \subseteq C$ for some $\epsilon > 0$.

Because $y \in \overline{C}, (\forall \epsilon > 0), y \in C + \epsilon B$. Hence,

$$\begin{aligned}
(1-\lambda)x + \lambda y + \epsilon B &\subseteq (1-\lambda)x + \lambda(C + \epsilon B) + \epsilon B \\
&= (1-\lambda)x + (1+\lambda)\epsilon B + \lambda C \\
&= (1-\lambda)\left(x + \frac{1+\lambda}{1-\lambda}\epsilon B\right) + \lambda C \\
&\subseteq (1-\lambda)C + \lambda C \qquad \text{for sufficiently small } \epsilon \\
&= C
\end{aligned}$$

Theorem 6.1: Let C be a convex set in $\mathbb{R}^n$. Then $(\forall x \in \text{ri}(C))(\forall y \in \overline{C})$, we have $[x, y) \subseteq \text{ri}(C)$.

> This is the relative interior version of Proposition L3-c.

Proof: for the case $\text{int}(C) \neq \emptyset$. The result follows from Proposition L3-b and Proposition L3-c.

In the case $\text{int}(C) = \emptyset$. In this case we must have $\dim C = m < n$. Let $L = \text{aff}(C) - \text{aff}(C)$. Then L is a linear subspace whose dimension is m.

We can find an isomorphism between L and $\mathbb{R}^m$ that preserves distance (and thus shape). From now on, we assume that $C \subseteq \mathbb{R}^m$, and the interior of C with respect to $\mathbb{R}^m$ is simply $\text{ri}(C)$ in $\mathbb{R}^n$. Now we can apply the first case. □

Theorem L3-d: Let C be a convex subset of $\mathbb{R}^n$. Then the following hold:

1. $\overline{C}$ is convex.
2. $\text{int}(C)$ is convex.
3. Suppose that $\text{int}(C) \neq \emptyset$. Then $\text{int}(C) = \text{int}(\overline{C})$ and $\overline{C} = \overline{\text{int}(C)}$

Proof: For point 1. Let $x, y \in \overline{C}$, and let $\lambda \in (0, 1)$. By closure property there exist sequences $(x_n), (y_n)$ in C such that $x_n \to x$ and $y_n \to y$. Consequently, $C \ni \lambda x_n + (1-\lambda)y_n \to \lambda x + (1-\lambda)y \implies \lambda x + (1-\lambda)y \in \overline{C}$. Hence $\overline{C}$ is convex.

For point 2. If $\text{int}(C) = \emptyset$, the result is vacuously true. Otherwise, use Proposition 3-2 with both $x, y \in \text{int}(C)$. And we arrive at $[x, y] \subseteq \text{int}(C)$.

For point 3. We first note that $C \subseteq \overline{C} \implies \text{int}(C) \subseteq \text{int}(\overline{C})$. Conversely, let $y \in \text{int}(\overline{C})$. Then by definition of interior, $\exists \epsilon > 0$ such that $B(y, \epsilon) \subseteq \overline{C}$. Now, let $x \in \text{int}(C)$, find $\lambda > 0$ such that $x \neq y$ and $y + \lambda(y - x) \in B(y, \epsilon) \subseteq \overline{C}$. Define $y' = y + \lambda(y - x)$. We apply **proposition L3-c**, with x, y', and we arrive at

$$y \in [x, y') \subseteq \text{int}(C)$$

Therefore $\text{int}(\overline{C}) \subseteq \text{int}(C)$ and we are done.

Moving on the the second identity. Clearly $\overline{\text{int}(C)} \subseteq \overline{C}$. The converse could be proven using a very similar technique as the first identity. Choose $y \in \overline{C}$ and $x \in \text{int}(C)$. $[x, y) \subseteq \text{int}(C)$ by proposition L3-c. Moreover, y is a limit point of $[x, y)$, hence $y \in \overline{\text{int}(C)}$.0 □

Theorem 6.2: Let C be a convex subset of $\mathbb{R}^n$. Then $\text{ri}(C)$ and $\overline{C}$ are convex subsets of $\mathbb{R}^n$. Moreover,

$$C \neq \emptyset \iff \text{ri}(C) \neq \emptyset$$

Separation Theorems

Let C_1, C_2 be subsets of $\mathbb{R}^n$. Then C_1, C_2 are **separated** if $\exists b \in \mathbb{R}^n \backslash \{0\}$ such that

$$\sup_{c_1 \in C_1} \langle c_1, b \rangle \leq \inf_{c_2 \in C_2} \langle c_2, b \rangle$$

C_1, C_2 are **strongly separated** if $\exists b \in \mathbb{R}^n \backslash \{0\}$ such that

$$\sup_{c_1 \in C_1} \langle c_1, b \rangle < \inf_{c_2 \in C_2} \langle c_2, b \rangle$$

We say that $x \in \mathbb{R}^n$ is (strongly) separated from $C \subseteq \mathbb{R}^n$ if the set $\{x\}$ is (strongly) separated from C.

Additionally, we say that C_1, C_2 are **properly separated** if $\exists b \neq 0$ such that

$$\inf_{c_1 \in C_1} \langle b, c_1 \rangle < \sup_{c_2 \in C_2} \langle b, c_2 \rangle \quad \text{and} \quad \sup_{c_1 \in C_1} \langle c_1, b \rangle \leq \inf_{c_2 \in C_2} \langle c_2, b \rangle$$

> Proper separation would allow at most one set to be in the separating hyperplane. Note that if both sets are in the separating hyperplane H (heck, if $C_1 = C_2 \subseteq H$), then they would be separated by definition.

Theorem L3-e: Let C be a nonempty, closed, convex subset of $\mathbb{R}^n$ and suppose that $x \notin C$. Then x is strongly separated from C.

Proof: we need to guarantee that $\exists b \in \mathbb{R}^n, b \neq 0$ such that

$$\sup\langle C, b \rangle < \inf\langle x, b \rangle = \langle x, b \rangle$$

Hence we only need to show $\exists b \neq 0$ such that $\sup\langle (C - x), b \rangle < 0$.

Set $b = x - P_C(x)$. (Note that $b \neq 0$ since $x \notin C$). Let $y = C$ and use projection theorem. We can derive to (exercise)

$$\sup_{y \in C} \langle y, b \rangle - \langle x, b \rangle < 0$$

And we're done. □

Corollary L3-f: Let C_1, C_2 be nonempty subsets of $\mathbb{R}^n$ such that $C_1 \cap C_2 = \emptyset$ and $C_1 - C_2$ is closed and convex. Then C_1, C_2 are strongly separated.

Proof: Observe that by definition, C_1, C_2 are strongly separated if and only if $C_1 - C_2$ are 0 are strongly separated. By definition of separation, $C_1 - C_2$ are 0 are strongly separated if and only if there exists $b \neq 0$ such that

$$\sup_{c_1 \in C_1, c_2 \in C_2} \langle c_1 - c_2, b \rangle < \inf\langle 0, b \rangle = 0$$
$$\iff \sup_{c_1 \in C_1} \langle c_1, b \rangle < -\sup_{c_2 \in C_2} \langle -c_2, b \rangle = \inf_{c_2 \in C_2} \langle c_2, b \rangle$$

END of observation. Note that $C_1 \cap C_2 = \emptyset \implies 0 \notin C_1 - C_2$ and combining with Theorem L3-e, we're done. □

Corollary L3-g: Let C_1, C_2 be nonempty, closed convex subset of $\mathbb{R}^n$ such that $C_1 \cap C_2 = \emptyset$ and C_2 is bounded. Then C_1 and C_2$ are strongly separated.

Proof: by Proposition L3-a, $C_1 - C_2$ is nonempty, closed and convex. The result then follows from Corollary L3-f.

Theorem L4-a: Suppose that C_1, C_2 are nonempty, closed convex subsets of $\mathbb{R}^n$ such that $C_1 \cap C_2 = \emptyset$. Then C_1, C_2 are separated.

> Note that the constraints are relaxed and we don't have strong separation anymore. According to professor: If we drop the condition of closedness, we can still conclude that C_1 and C_2 are separated. Refer to Textbook Theorem 11.3.
>
> We work with closed sets in this course because some proofs simplify. Remember closeness is critical for the existence of the projection onto the set. Projections onto closed convex sets will be key ingredients in the algorithms we study.

Proof: Set $(\forall n \in \mathbb{N}), D_n = C_2 \cap B(0, n)$. Observe that $C_1 \cap D_n = \emptyset$ and D_n is bounded. Duh.

Apply Corollary L3-g. There exists a hyperplane that strongly separates C_1 and D_n. Equivalently, $\forall n \in \mathbb{N}, \exists u_n \in \mathbb{R}^m$ such that $\|u_n\| = 1$ and $\sup\langle C_1, u_n \rangle < \inf\langle D_n, u_n \rangle$.

Because $(u_n)_{n \in \mathbb{N}}$ is bounded, there exists a convergent subsequence $(u_{k_n})_{n \in \mathbb{N}}$ of $(u_n)_{n \in \mathbb{N}}$ such that say $u_{k_n} \to u$ and $\|u\| = 1$.

Now let $x \in C_1, y \in C_2$. Then eventually $\exists l \in \mathbb{N}$ such that for all $k_n \geq l$, we have $y \in B(0, k_n)$, hence eventually $y \in D_{k_n}$, and by

$$\langle x, u_{k_n} \rangle < \langle y, u_{k_n} \rangle$$

Taking the limit $k_n \to \infty$, we learn that $\langle x, u \rangle \leq \langle y, u \rangle$. □

Cone: Let C be a subset of $\mathbb{R}^n$. Then

1. C is a **cone** if $\forall \lambda \in (0, +\infty), C = \lambda C$
2. **Conical Hull** of C, denoted by $\text{cone}(C)$, is the intersection of all the cones of $\mathbb{R}^n$ containing C. It is the smallest cone in $\mathbb{R}^n$ containing C.
3. **Closed conical hull** of C, denoted by $\overline{\text{cone}}(C)$ is the smallest closed cone in $\mathbb{R}^n$ containing C.

> Note that in the context of this course, a cone is not automatically convex. "Convex cone" will be explicitly stated.

Proposition L4-b: Let C be a subset of $\mathbb{R}^n$. Then the following hold:

1. $\overline{\text{cone}(C)} = \{\lambda c : c \in C, \lambda \in [0, \infty)\}$
2. $\overline{\text{cone}(C)} = \overline{\text{cone}}(C)$
3. $\text{cone}(\text{conv}(C)) = \text{conv}(\text{cone}(C))$
4. $\overline{\text{cone}}(\text{conv}(C)) = \overline{\text{conv}}(\text{cone}(C))$

Proof: If $C = \emptyset$ then the conclusion is obvious. Now suppose that $C \neq \emptyset$.

For Part 1, $D = \{\lambda c : c \in C, \lambda \in [0, \infty)\}$. Note that $C \subseteq D$ and D is a cone. Therefore $\text{cone}(C) \subseteq D$. Conversely, let $y \in D$. Then $\exists \lambda > 0, c \in C$ such that $y = \lambda c$. This means y is in all the cones that contain C and therefore $y \in \text{cone}(C)$. And we're done.

For Part 2. Observe that $\overline{\text{cone}}(C)$ is closed cone. Clearly, $C \subseteq \overline{\text{cone}}(C)$. Hence

$$\overline{\text{cone}(C)} \subseteq \overline{\text{cone}(\overline{\text{cone}}(C))} = \overline{\text{cone}}(C)$$

Conversely, since $\overline{\text{cone}(C)}$ is a cone, it follows that $\overline{\text{cone}}(C) \subseteq \overline{\text{cone}(C)}$. And we're done.

For part 3. Hint: Caratheodory's Theorem + Brute Force derivation.

For part 4. This is a direct consequence of part 2,3. And the fact $\overline{\text{conv}(X)} = \overline{\text{conv}}(X)$ (follows from Theorem L3-d). □

Lemma L4-c: Let C be a convex subset of $\mathbb{R}^n$ such that $\text{int}(C) \neq \emptyset$ and $0 \in C$. Then the following are the equivalent:

1. $0 \in \text{int}(C)$
2. $\text{cone}(C) = \mathbb{R}^n$
3. $\overline{\text{cone}}(C) = \mathbb{R}^n$

Proof: $(1. \implies 2.)$, note $\mathbb{R}^n = \text{cone}(B(0, \epsilon)) \subseteq \text{cone}(C) \subseteq \mathbb{R}^n$.

$(2. \implies 3.)$, by Proposition L4-b, the following holds: $\mathbb{R}^n = \text{cone}(C) \subseteq \overline{\text{cone}(C)} = \overline{\text{cone}}(C) \subseteq \mathbb{R}^n$.

$(3. \implies 1.)$, since C is a convex set, we have $C = \text{conv}(C)$. Apply theorem L4-b, we have $\text{cone}(\text{conv}(C)) = \text{cone}(C) = \text{conv}(\text{cone}(C))$. Moreover, $\emptyset \neq \text{int}(C) \subseteq \text{int}(\text{conv}(C))$. Therefore, $\text{cone}(C)$ is a convex set, and $\text{int}(\text{cone}(C)) \neq \emptyset$.

by Theorem L3-d, $\text{int}(\text{cone}(C)) = \text{int}(\overline{\text{cone}(C)}) = \text{int}(\overline{\text{cone}}(C))$.

> Auxiliary Fact (Exercise): Let C be a convex subset of $\mathbb{R}^n$ such that $\text{int}(C) \neq \emptyset$ and $0 \in C$. Then
>
> $$\text{int}(\text{cone}(C)) = \text{cone}(\text{int}(C))$$

$$\begin{aligned}
\mathbb{R}^n &= \text{int}(\mathbb{R}^n) \\
&= \text{int}(\overline{\text{cone}}(C)) \\
&= \text{int}(\text{cone}(C)) \\
&= \text{cone}(\text{int}(C)) \\
\implies \quad & 0 \in \text{cone}(\text{int}(C)) \\
\implies \quad & \lambda \cdot 0 \in \text{int}(C), \quad \text{for some } \lambda > 0 \\
\implies \quad & 0 \in \text{int}(C)
\end{aligned}$$

And we're done. □

Tangent and Normal Cones

Let C be a nonempty convex subset of $\mathbb{R}^n$ and let $x \in \mathbb{R}^n$. The **tangent cone** to C at x is:

$$T_C(x) = \begin{cases} \overline{\text{cone}}(C - \{x\}), & x \in C \\ \emptyset, & x \notin C \end{cases}$$

And the normal cone of C at x is

$$N_C(x) = \begin{cases} \{u \in \mathbb{R}^n : \sup_{c \in C} \langle c - x, u \rangle \leq 0\}, & x \in C \\ \emptyset, & x \notin C \end{cases}$$

> Note that this definition of normal cone is the negative of the normal cone in CO 255. So that for this course, tangent cone and normal cone are facing away from each other.

Lemma L4-d: Let C be a non-empty closed convex subset of $\mathbb{R}^n$ and let $x \in C$. Then $n \in N_C(x)$ if and only if $\forall t \in T_C(x)$ we have $\langle n, t \rangle \leq 0$.

Proof: $(\Rightarrow)$ Let $n \in N_C(x)$, and let $t \in T_C(x)$. Since $T_C(x) = \overline{\text{cone}}(C - \{x\})$, there exists $\lambda_k > 0$ and sequence $(t_k)_{k \in \mathbb{N}}$ in $\mathbb{R}^n$ such that $(\forall k \in \mathbb{N})$, $x + \lambda_k t_k \in C$ and $t_k \to t$.

By definition of normal cone, we know $\langle n, \lambda_k t_k \rangle \leq 0$, and then $\langle n, t_k \rangle \leq 0$. As $k \to +\infty$ we have $\langle n, t \rangle \leq 0$.

$(\Leftarrow)$: suppose that $\forall t \in T_C(x)$ we have $\langle n, t \rangle \leq 0$. Let $y \in C$, then by definition of tangent cone $y - x \in T_C(x)$. Therefore, $\langle n, y - x \rangle \leq 0 \implies n \in N_C(x)$, by definition of normal cone. □

Lemma L4-e: Let C be a on-empty closed convex subset of $\mathbb{R}^n$ and let $x \in C$. Then $n \in T_C(x)$ if and only if $(\forall n \in N_C(x)), \langle n, t \rangle \leq 0$.

> Note this Lemma is not used, (should've been used for the last part of proof for the next theorem).

Theorem L4-f: Let C be a convex subset of $\mathbb{R}^n$ such that $\text{int}(C) \neq \emptyset$ and let $x \in C$. Then the following is equivalent:

1. $x \in \text{int}(C)$
2. $T_C(x) = \mathbb{R}^n$
3. $N_C(x) = \{0\}$

Proof: $(1. \iff 2.)$ using lemma L4-c we have $x \in \text{int}(C) \iff 0 \in \text{int}(C - \{x\}) \iff \overline{\text{cone}}(C - x) = \mathbb{R}^n \iff T_C(x) = \mathbb{R}^n$.

$(2. \iff 3.)$: using lemma L4-d, let $n \in N_C(x)$, then for all $t \in T_C(x) = \mathbb{R}^n$, $\langle n, t \rangle \leq 0$. This implies that $\langle n, n \rangle \leq 0 \iff n = 0$. This gives $N_C(x) \subseteq \{0\}$. Clearly, $\{0\} \subseteq N_C(x)$. We arrive at $N_C(x) = \{0\}$.

Conversely, let $N_C(x) = \{0\}$. Set $K = T_C(x)$, recall that K is closed **convex** cone with $0 \in K$. Let $x \in \mathbb{R}^n$, (we want to show that $x \in K$,) and set $p = P_K(x)$. Then by the projection theorem $(\forall y \in K), \langle x - p, y - p \rangle \leq 0$, (I.). In particular:

$$\begin{cases} \langle x - p, -p \rangle \leq 0 & (\text{setting y=0}) \\ \langle x - p, p \rangle \leq 0 & (\text{setting y=2p}) \end{cases} \implies \langle x - p, p \rangle = 0$$

Hence, (I.) becomes $(\forall y \in K) \langle x - p, y \rangle \leq 0$. It follows from the lemma in that $x - p \in N_C(x) = \{0\}$. As a result, $x = p = P_K(x) \in K$ by definition of projection. □

More on Convex

Epigraph: let $f : \mathbb{R}^n \longrightarrow [-\infty, +\infty]$, note that $[-\infty, +\infty] = \mathbb{R} \cup \{\pm\infty\}$ is the **extended real line**. The epigraph of f is

$$\text{epi}(f) = \{(x, \alpha) : f(x) \leq \alpha\} \subseteq \mathbb{R}^n \times \mathbb{R}$$

Domain: Let $f : \mathbb{R}^n \longrightarrow [-\infty, +\infty]$. Then $\text{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}$.

> Note that this domain is different from the domain in calculus (which says domain is where the function is defined).

f is **proper** if $\text{dom}(f) \neq \emptyset$ and $\forall x \in \mathbb{R}^n, f(x) > -\infty$.

Examples:

- Let $f : \mathbb{R}^m \longrightarrow (-\infty, +\infty)$ be continuous. Then f is proper.
- Let C be a subset of $\mathbb{R}^m$. The **indicator function** of C at $x \in \mathbb{R}^m$ (see textbook p.28) is

$$\delta_C(x) = \begin{cases} 0, & x \in C \\ +\infty & \text{otherwise} \end{cases}$$

  Clearly, δ_C is proper whenever $C \neq \emptyset$. (Also note that the domain of the indicator function is always C)

Properties of f:

- f is **lower semicontinuous** (**l.s.c.**) if $\text{epi}(f)$ is closed
- f is **convex** if $\text{epi}(f)$ is convex

Proposition L5-a: Let $f : \mathbb{R}^m \longrightarrow [-\infty, +\infty]$ be convex, then $\text{dom}(f) = \{x \in \mathbb{R}^n : f(x) < +\infty\}$ is convex.

> A Fact that we use in proof: Let C be a convex subset of $\mathbb{R}^n$ and let $A : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. If C is a convex subset of $\mathbb{R}^n$ then $A(C)$ is a convex subset of $\mathbb{R}^m$.

Proof: Recall that $\text{epi}(f) = \{(x, \alpha) : f(x) \leq \alpha\} \subseteq \mathbb{R}^{n+1}$. Consider the linear map (transformation) $L : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^n$: $(x, \alpha) \longrightarrow x$. Then $\text{dom} = L(\text{epi}(f))$, and the conclusion follows in view of the above fact.

(Also note that the definition of epigraph requires α to be on the real line, which always satisfy $\alpha < +\infty$.) □

Theorem L5-b: Let $f : \mathbb{R}^m \longrightarrow [-\infty, +\infty]$. Then f is convex if and only if $(\forall x \in \text{dom}(f))\ (\forall y \in \text{dom}(f))\ (\forall \lambda \in (0, 1))$, we have

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

Proof: Observe that $f = +\infty \iff \text{epi}(f) = \emptyset \iff \text{dom}(f) = \emptyset$ and the conclusion follows. Now suppose $\text{dom}(f) \neq \emptyset$.

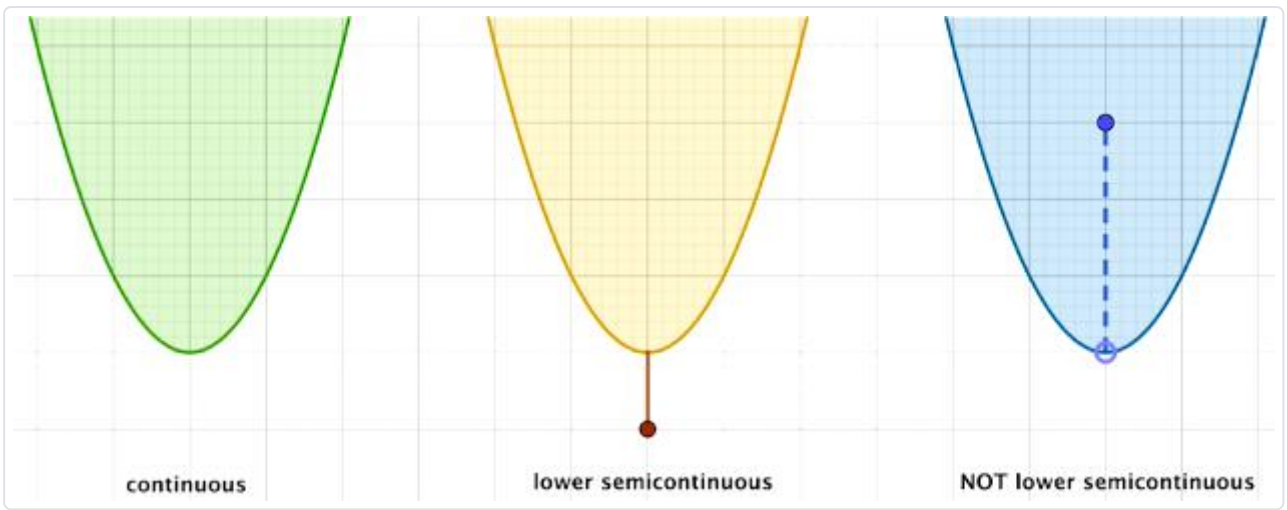
$(\Rightarrow)$: let $x, y \in \text{dom}(f)$. The result follows from the definition of convexity and epigraph.

$(\Leftarrow)$: Let $(x, \alpha) \in \text{epi}(f), (y, \beta) \in \text{epi}(f), \lambda \in (0, 1)$. By definition of epigraph: $f(x) \leq \alpha, f(y) \leq \beta$. Now

$$\begin{aligned} f(\lambda x + (1 - \lambda)y) &\leq \lambda f(x) + (1 - \lambda) f(y) \\ &\leq \lambda\alpha + (1 - \lambda)\beta \end{aligned}$$

Consequently, $(\lambda x + (1 - \lambda)y, \lambda\alpha + (1 - \lambda)\beta) = \lambda(x, \alpha) + (1 - \lambda)(y, \beta) \in \text{epi}(f)$. That is, $\text{epi}(f)$ is convex, i.e. f is convex.

Sequential Characterization of Lower Semicontinuity (alternative definition): Let $f : \mathbb{R}^m \to [-\infty, +\infty]$ and let $x \in \mathbb{R}^n$. Then f is lower semicontinuous (l.s.c.) at x if for every sequence $(x_n)_{n \in \mathbb{N}}$ in $\mathbb{R}^m$, $x_n \to x$ implies $f(x) \leq \liminf f(x_n)$. Moreover, f is l.s.c. if f is l.s.c. at every point in $\mathbb{R}^m$.



Remarks:

1. If f is continuous then f is l.s.c.
2. One can show that the equivalence of the definitions of l.s.c. However, we will omit the proof.

Theorem L5-c: Let $C \subseteq \mathbb{R}^m$. Then the following hold (recall the indicator function).

1. $C \neq \emptyset \iff \delta_C$ is proper.
2. C is convex $\iff \delta_C$ is convex.
3. C is closed $\iff \delta_C$ is l.s.c.

Proof for part 3: Observe that $C = \emptyset \iff \mathrm{epi}(\delta_C) = \emptyset$ which is closed. Now suppose $C \neq \emptyset$:

$(\Rightarrow)$: we want to show that the $\mathrm{epi}(\delta_C)$ is closed. Take a sequence $((x_n, \alpha_n))_{n\in\mathbb{N}}$ be a sequence in $\mathrm{epi}(\delta_C)$ such that $(x_n, \alpha_n) \to (x, \alpha)$.

Observe that $(x_n)_{n\in\mathbb{N}}$ is a sequence in C, $x_n \to x$. We also have $x \in C$ due to closedness. Then, $(\forall n \in \mathbb{N}), 0 = \delta_C(x_n) \leq \alpha_n \implies 0 = \delta_C(x) \leq \alpha \implies (x, \alpha) \in \mathrm{epi}(\delta_C)$, and we're done.

$(\Leftarrow)$: Conversely, suppose that δ_C is l.s.c. Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in C such that $x_n \to x$. We want to show that $x \in C$, and it is sufficient to show that $\delta_C(x) = 0$. By sequential characteristization of l.s.c. we have

$$0 \leq \delta_C(x) \leq \liminf \delta_C(x_n) = 0$$

Hence $\delta_C(x) = 0$. □

Proposition L5-d: Let I be an indexed set and let $(f_i)_{i\in I}$ be a family of l.s.c. convex functions on $\mathbb{R}^n$. Then $\sup\limits_{i\in I} f_i$ is convex and l.s.c.

Proof: Set $F = \sup\limits_{i\in I} f_i$. We claim that $\mathrm{epi}(F) = \bigcap\limits_{i\in I} \mathrm{epi}(f_i)$. Indeed, let $(x, \alpha) \in \mathbb{R}^m \times \mathbb{R}$. Then,

$$\begin{aligned}(x, \alpha) \in \mathrm{epi}(F) &\iff \sup_{i\in I} f_i(x) \leq \alpha \\ &\iff (\forall i \in I)(x, \alpha) \in \mathrm{epi}(f_i) \\ &\iff (x, \alpha) \in \bigcap_{i\in I} \mathrm{epi}(f_i)\end{aligned}$$

This proves the claim. Now since the intersection of convex sets is convex, and the intersection of l.s.c. sets is l.s.c. We arrive at the result. □

Support Function: Let C be a subset of $\mathbb{R}^m$. The support function of C is

$$\begin{aligned}\sigma_C : \mathbb{R}^m &\longrightarrow [-\infty, +\infty] \\ u &\longmapsto \sup_{c\in C} \langle c, u\rangle\end{aligned}$$

Proposition L5-e: Let C be a nonempty subset of $\mathbb{R}^n$. Then σ_C is convex, l.s.c. and proper.

Proof: Let $c \in C$ and set $f_c : \mathbb{R}^m \longrightarrow \mathbb{R} : x \longmapsto \langle x, c\rangle$. Then f_c is proper, (l.s).c. and convex. (In fact, f_c is linear). Note that $\sigma_C = \sup\limits_{c\in C} f_C$. Now by Proposition L5-d, we have σ_C is convex and l.s.c.

To demonstrate that σ_C is proper, since $C \neq \emptyset, \sigma_C(0) = \sup\limits_{c\in C}\langle 0, c\rangle = 0 < +\infty$. Hence, $0 \in \mathrm{dom}(\sigma_C) \neq \emptyset$. Moreover, let $\bar{c} \in C$. Then $(\forall u \in \mathbb{R}^m) \quad \sigma_C(u) = \sup\limits_{c\in C}\langle u, c\rangle \geq \langle u, \bar{c}\rangle > -\infty$. Hence σ_C is proper. □

Example: Let $C = [a, b] \subseteq \mathbb{R}^+ \cup \{0\}$. Then $(\forall x \in \mathbb{R}), \sigma_C(x) = \sup_{c \in [a,b]} cx = \begin{cases} bx, & x \geq 0 \\ ax, & x < 0 \end{cases}$

Example: Let $C = [0, +\infty) \subseteq \mathbb{R}$. We examine two cases:

- Case 1: $x \leq 0$. Then $\sigma_C(x) = 0$.
- Case 2: $x > 0$. Then $\sigma_C(x) = +\infty$.

Hence $\mathrm{dom}(\sigma_C) = (-\infty, 0]$.

Further notions of convexity: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper. Then f is

- **Strictly Convex** if $(\forall x, y \in \mathrm{dom}(f))(\forall \lambda \in (0,1)) \quad x \neq y \implies f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)$.
- **Strongly Convex** with constant β, if for some $\beta > 0$ we have $(\forall x, y \in \mathrm{dom}(f))(\forall \lambda \in (0,1)) \quad f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) - \frac{\beta}{2}\lambda(1-\lambda)\|x - y\|^2$.

Clearly, $\text{strong convexity} \implies \text{strict convexity} \implies \text{convexity}$.

Proposition L6-a: Let I be a finite indexed set, let $(f_i)_{i \in I}$ be a family of convex functions from $\mathbb{R}^m$ to $[-\infty, +\infty]$. Then $\sum_{i \in I} f_i$ is convex.

Proposition L6-b: Let f be convex and l.s.c. and let $\lambda > 0$. Then λf is convex and l.s.c.

Minimizers of functions: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper and let $x \in \mathbb{R}^m$. Then x is a (global) minimizer of f if $f(x) = \min(f(\mathbb{R}^m) \in \mathbb{R})$. Throughout the course, we will use $\mathrm{Argmin}(f)$ to denote the set of minimizers of f.

Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper and let $\bar{x} \in \mathbb{R}^m$. Then

- $\bar{x}$ is a local minimum of f if $\exists \delta > 0$ such that $\|x - \bar{x}\| < \delta \implies f(\bar{x}) \leq f(x)$.
- $\bar{x}$ is a global minimum of f if $(\forall x \in \mathrm{dom}(f)) \quad f(\bar{x}) \leq f(x)$.

Analogously, we can define local/global max.

Proposition L6-c: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper and convex. Then every local minimizer of f is a global minimizer.

Proof: Let x be a local minimizer of f. Then $\exists \rho > 0$ such that $f(x) = \min(f(B(x, \rho)))$. We want to show that x is a global minimizer of f, i.e. $(\forall y \in \mathrm{dom}(f)) \quad f(x) \leq f(y)$.

Let $y \in \mathrm{dom}(f)$ and observe that if $y \in B(x, \rho)$, i.e. $\|x - y\| \leq \rho$, then $f(x) \leq f(y)$. Now suppose that $y \in \mathrm{dom}(f) \backslash B(x, \rho)$. Observe that $\lambda := 1 - \frac{\rho}{\|x - y\|} \in (0, 1)$. Set $z = \lambda x + (1 - \lambda) y \in \mathrm{dom}(f)$, which is true because $\mathrm{dom}(f)$ is convex. $\|z - x\| = \rho$ and as such $z \in B(x, \rho)$.

Moreover, because f is convex, it follows from **Jensen's inequality** that

$$\begin{aligned} f(x) &\leq f(z) \\ &= f(\lambda x + (1-\lambda) y) \\ &\leq \lambda f(x) + (1-\lambda) f(y) \\ \implies (1-\lambda) f(x) &\leq (1-\lambda) f(y) \\ \implies f(x) &\leq f(y) \end{aligned}$$

And we're done. □

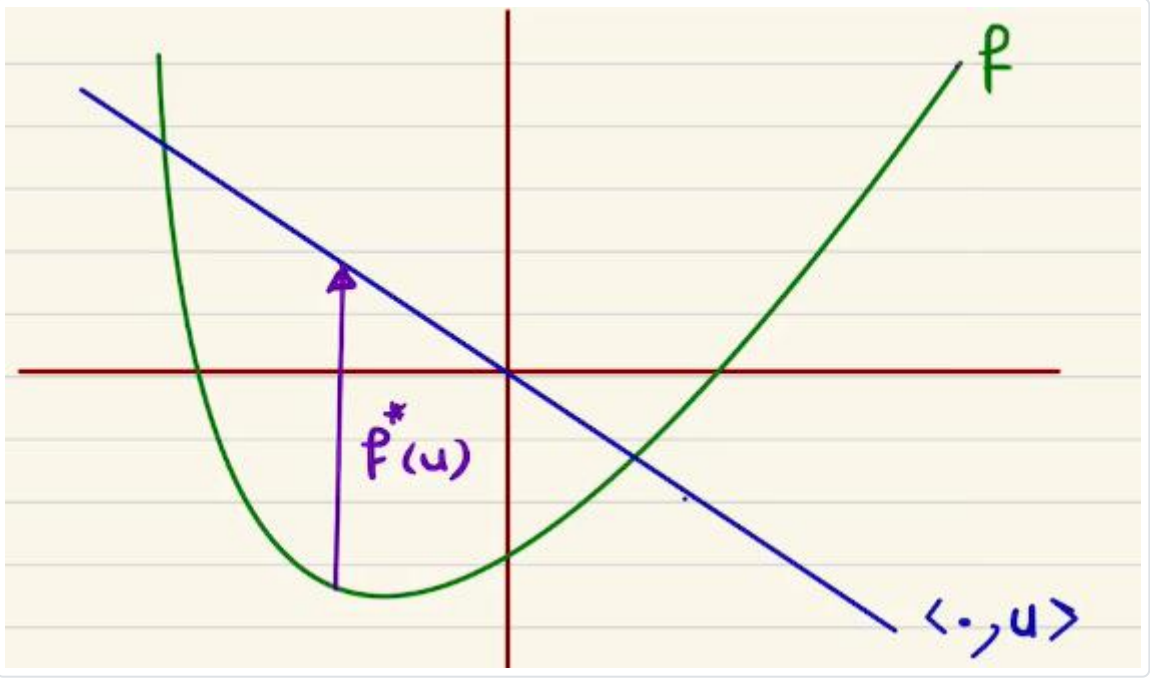
Proposition L6-d: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper and convex, and let C be a subset of $\mathbb{R}^m$. Suppose that x is a minimizer of f over C such that $x \in \mathrm{int}(C)$. Then x is a minimizer of f.

Proof: since $x \in \mathrm{int}(C), \exists \epsilon > 0$ such that $B(x, \epsilon) \subseteq C$. The result then follows from L6-c. □

Conjugates of Convex Functions: Let $f : \mathbb{R}^m \to [-\infty, +\infty]$. The **Fenchel-Legendre Transformation** or **Convex Conjugate** of f is

$$\begin{aligned} f^* : \mathbb{R}^m &\longrightarrow [-\infty, +\infty] \\ u &\longmapsto \sup_{x \in \mathbb{R}^m} \left(\langle x, u \rangle - f(x) \right) \end{aligned}$$

To visualize convex conjugate: basically the input u determins the slope of function $\langle \cdot, u \rangle$ and convex functions are guaranteed to have upwards "curvature." And so the supremum is well defined.



Proposition L6-e: Let $f : \mathbb{R}^m \to [-\infty, \infty]$. Then f^* is convex and l.s.c.

Proof: Observe that if $f = +\infty \iff \mathrm{dom}(f) = \emptyset$. Hence, $(\forall u \in \mathbb{R}^m) \quad f^*(u) = \sup\limits_{x \in \mathbb{R}^m} (\langle x, u \rangle - f(x)) = \sup\limits_{x \in \mathrm{dom}(f)} (\langle x, u \rangle - f(x)) = -\infty$ because the supremum of empty set is $-\infty$ (by convention, and it makes sense).

Now suppose that $f \not\equiv +\infty$, define $g_x(u) = \langle x, u \rangle - f(x)$ for some $x \in \mathbb{R}^m$. Note that $f^* = \sup\limits_{x \in \mathbb{R}^m} g_x$, and that g_x is an affine function for a given x, (indeed, $f(x)$ is considered constant in this context). Since affineness implies convexity and l.s.c. property, the result follows from proposition L5-d.

Example L6-7: Let $p > 1$ and set $q = \dfrac{p}{p-1}$. Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto \dfrac{|x|^p}{p}$. Then $f^* : \mathbb{R} \to \mathbb{R} : u \mapsto \dfrac{|u|^q}{q}$.

Proof: For some $u \in \mathbb{R}$, define $g(x) := xu - f(x) = xu - \frac{|x|^p}{p}$. Note that $f^*(u) = \sup\limits_{x \in \mathbb{R}} g(x)$, and observe that f, g, are both differentiable on $\mathbb{R}$ (verification as exercise).

$$g'(x) = u - \begin{cases} x^{p-1}, & x \geq 0 \\ -(-x)^{p-1} = -(|x|)^{p-1}, & x < 0 \end{cases}$$

$g(x)$ only has a unique local minimum (which is also the global minimum). It could be visualized because it is the difference between a line passing through origin, and the absolute value of a polynomial function (resembling quadratic). Setting $g'(x) = 0$, we get $|x| = |u|^{\frac{1}{p-1}}$ and $\mathrm{sgn}(x) = \mathrm{sgn}(u)$, for the given u.

As a result, substitute $|x|$ at global maximum into $g(x)$, and we will arrive at the result. □

Example L6-8: Let $f : \mathbb{R} \to \mathbb{R}, f(x) = e^x$. Then $f^*(u) = \begin{cases} u\ln(u) - u, & u > 0 \\ 0, & u = 0 \\ +\infty, & u < 0 \end{cases}$

Note L6-f: Let C be a subset of $\mathbb{R}^m$. Then $\delta_C^* = \sigma_C$.

Proof: Indeed, $\delta_C^*(u) = \sup\limits_{y \in C}(\langle x, y \rangle - \delta_C(y)) = \sup\limits_{y \in C} \langle x, y \rangle$. □

Subgradient Calculus

Subdifferential Operator: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper. The subdifferential of f is the set-valued operator

$$\begin{aligned} \partial f : \mathbb{R}^m \rightrightarrows \mathbb{R}^m \qquad & \text{denote set-valued operator} \\ : x \mapsto \{u \in \mathbb{R}^m : (\forall y \in \mathbb{R}^m)\ f(y) \geq f(x) + \langle u, y - x \rangle\} & \end{aligned}$$

Let $x \in \mathbb{R}^m$. Then f is subdifferentiable at x if $\partial f(x) \neq \emptyset$.

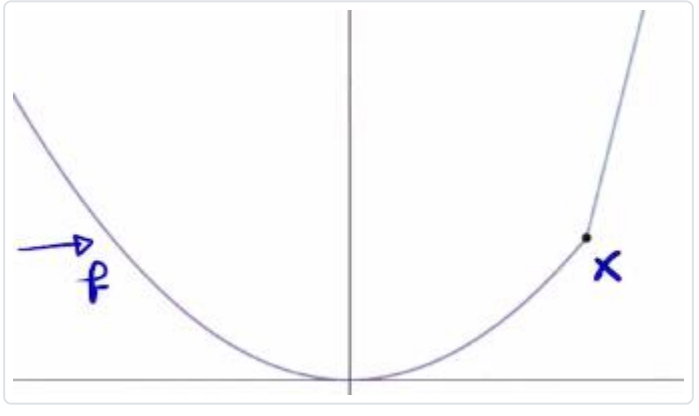
Fermat's Theorem L6-g: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper. Then

$$\mathrm{Argmin}(f) = \{x \in \mathbb{R}^m : 0 \in \partial f(x)\} =: \mathrm{zer}(\partial f)$$

> Note that this is analogous to $\nabla f = 0$, but this is the price we pay for the lack of differentiability. You can think of this as a generalized version of gradient.

Proof: follow from definition of minimizer and then subdifferential operator. □

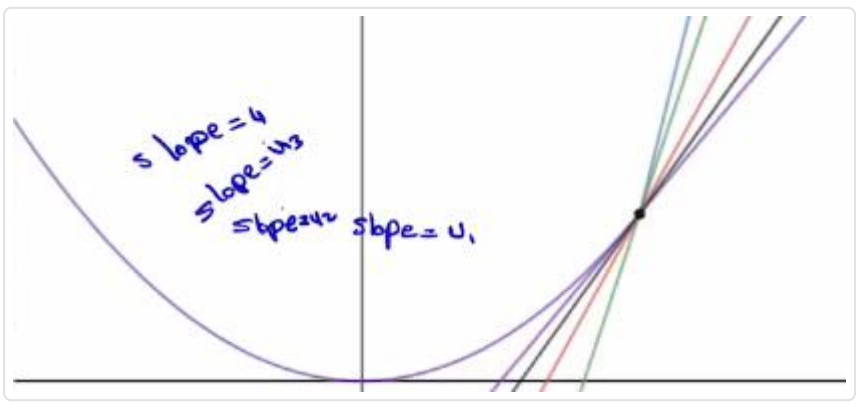
Consider the function:



u is a **subgradient** of f at x, if for all $y \in \mathbb{R}^m$ we have

$$f(y) \geq f(x) + \langle u, y - x \rangle =: h(y)$$

$h(y)$ is an **affine function** that is a **lower bound** of $f(y)$, with slope = u. And note $h(x) = f(x)$, i.e. h coincide with f as well.



> Example: let $f : \mathbb{R} \to \mathbb{R} : x \mapsto |x|$, then $\partial f(x) = \begin{cases} \{-1\}, & x < 0 \\ [-1, 1], & x = 0 \\ \{1\}, & x > 0 \end{cases}$

Lemma L6-h: $f : \mathbb{R}^m \to (-\infty, \infty]$ is proper $\implies \mathrm{dom}(\partial f) \subseteq \mathrm{dom}(f)$.

> Note that $\mathrm{dom}(\partial f) = \{x : \partial f(x) \neq \emptyset\}$

Proof: Indeed, $f(x) = +\infty \implies \partial f(x) = \emptyset$.

Now, we use contrapositive: $x \notin \mathrm{dom}(f) \implies x \notin \mathrm{dom}(\partial f)$.

Note L6-i: Let C be a convex closed nonempty subset of $\mathbb{R}^m$, let $x \in \mathbb{R}^m$. Then

$$\partial \delta_C(x) = N_C(x)$$

Proof: Indeed, let $u \in \mathbb{R}^m$ and let $x \in C$, recall $\mathrm{dom}(\partial f) \subseteq \mathrm{dom}(f) = C$. Then

$$\begin{aligned} & u \in \partial \delta_C(x) \\ \iff & (\forall y \in \mathbb{R}^m)\ \delta_C(y) \geq \delta_C(x) + \langle u, y - x \rangle \\ \iff & (\forall y \in C)\ \delta_C(y) \geq \delta c(x) + \langle u, y - x \rangle \\ \iff & (\forall y \in C) 0 \geq \langle u, y - x \rangle \\ \iff & u \in N_C(x) \end{aligned}$$

This is actually easily visualizable. □

Calculus of Subdifferentials

From calculus. Let $f, g : \mathbb{R}^m \to (-\infty, +\infty]$ be proper and let $x \in \mathbb{R}^m$. Suppose that f, g are differentiable at x. Then

$$\nabla(f + g)(x) = \nabla f(x) + \nabla g(x)$$

Question: Let $f, g : \mathbb{R}^m \to (-\infty, +\infty]$ be proper, convex and l.s.c. and let $x \in \mathbb{R}^m$. Suppose that f, g are subdifferentiable at x. Then does the following hold?

$$\partial(f+g)(x) = \partial f(x) + \partial g(x)$$

Fact L7-a: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, lsc. and proper. Then

$$\emptyset \neq \operatorname{ri}(\operatorname{dom}(f)) \subseteq \operatorname{dom}(\partial f)$$

This guarantees the domain of the subdifferential is not empty (since domain is convex, and the relative interior of a nonempty convex set is nonempty, fact). In particular

$$\begin{aligned} \operatorname{ri}(\operatorname{dom}(f)) &= \operatorname{ri}(\operatorname{dom}(\partial f)) \\ \overline{\operatorname{dom}(f)} &= \overline{\operatorname{dom}(\partial f)} \end{aligned}$$

Theorem 11.3: Let C_1, C_2 be nonempty convex subsets of $\mathbb{R}^m$. Then C_1, C_2 are properly separated if and only if

$$\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$$

Corollary 6.6.2: Let C_1, C_2 be convex subsets of $\mathbb{R}^m$. Then $\operatorname{ri}(C_1 + C_2) = \operatorname{ri}(C_1) + \operatorname{ri}(C_2)$. Let $\lambda \in \mathbb{R}$. Then $\operatorname{ri}(\lambda C) = \lambda \operatorname{ri}(C)$.

Fact L7-b, (textbook top of page 49): Let $C_1 \subseteq \mathbb{R}^m, C_2 \subseteq \mathbb{R}^p$ be convex. Then $\operatorname{ri}(C_1 \oplus C_2) = \operatorname{ri}(C_1) \oplus \operatorname{ri}(C_2)$. Where $C_1 \oplus C_2 \cong C_1 \times C_2 = \{(c_1, c_2) : c_1 \in C_1, c_2 \in C_2\}$.

Theorem L7-c: Let C_1, C_2 be convex subsets of $\mathbb{R}^m$ such that $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) \neq \emptyset$. Let $x \in C_1 \cap C_2$. Then

$$N_{C_1 \cap C_2}(x) = N_{C_1}(x) + N_{C_2}(x)$$

> Note, this theorem is equivalent to Theorem 23-6 from CO 255. It is *the* big one.

Proof for $(\subseteq)$: Let $x \in C_1 \cap C_2$ and let $n \in N_{C_1 \cap C_2}(x)$. Then $(\forall y \in C_1 \cap C_2)$ we have $\langle n, y - x \rangle \leq 0$. Set

$$\begin{cases} E_1 = \operatorname{epi}(\delta_{C_1}) = C_1 \times [0, +\infty) \subseteq \mathbb{R}^m \times \mathbb{R} \\ E_2 = \{(y, \alpha) : y \in C_2, \alpha \leq \langle n, y - x \rangle\} \subseteq \mathbb{R}^m \times \mathbb{R} \end{cases}$$

Using Fact L7-b, applied with C_2 replaced by $[0, +\infty) \subseteq \mathbb{R}$, we learn that

$$\operatorname{ri}(E_1) = \operatorname{ri}(C_1) \times (0, +\infty)$$

We now claim that

$$\operatorname{ri}(E_1) \cap \operatorname{ri}(E_2) = \emptyset \tag{I.}$$

Indeed, suppose for a contradiction that $\exists (z, \alpha) \in \operatorname{ri}(E_1) \cap \operatorname{ri}(E_2)$. Then $0 < \alpha < \langle n, z - x \rangle \leq 0$, which is impossible.

Applying Theorem 11.3 with C_i replaced by E_i, we have E_1, E_2 are properly separated, which means that $\exists (b, \gamma) \in \mathbb{R}^m \times (\mathbb{R} \backslash \{0\})$ such that $(\forall (x, \alpha) \in E_1)(\forall (y, \beta) \in E_2)$ we have

$$\langle (x, \alpha), (b, \gamma) \rangle \leq \langle (y, \beta), (b, \gamma) \rangle \tag{II.}$$

Moreover, by proper separation, $(\exists (\bar{x}, \bar{\alpha}) \in E_1)(\exists (\bar{y}, \bar{\beta}) \in E_2)$ such that

$$\langle \bar{x}, b \rangle + \bar{\alpha}\gamma < \langle \bar{y}, b \rangle + \bar{\beta}\gamma \tag{III.}$$

We claim that $\gamma < 0$. Indeed, observe that $(x, 1) \in E_1, (x, 0) \in E_2$, combining with (II.) we have $\langle x, b \rangle + \gamma \leq \langle x, b \rangle \implies \gamma \leq 0$. Next, we show that $\gamma \neq 0$. Suppose for a contradiction that $\gamma = 0$. Observe that this implies that from (II.) and (III.), $\exists b \neq 0$ such that $\langle x, b \rangle \leq \langle y, b \rangle$ and $\langle \bar{x}, b \rangle < \langle \bar{y}, b \rangle$ and thus C_1, C_2 are properly separated. By Theorem 11.3, we get $\operatorname{ri}(C_1) \cap \operatorname{ri}(C_2) = \emptyset$ which contradicts our assumption in the theorem. Altogether, $\gamma < 0$.

Note that $n = -\dfrac{b}{\gamma} + n + \dfrac{b}{\gamma}$, we want to show that

$$-\frac{b}{\gamma} \in N_{C_1}(x) \qquad n + \frac{b}{\gamma} \in N_{C_2}(x)$$

Indeed, $(\forall y \in C_1)$ we have $(y, 0) \in E_1, (x, 0) \in E_2$. Substitute this into (II.), we get

$$
\begin{aligned}
&(\forall y \in C_1)\ \langle b, y\rangle \leq \langle b, x\rangle \\
\iff &(\forall y \in C_1)\ \langle b, y-x\rangle \leq 0 \\
\iff &b \in N_{C_1}(x)
\end{aligned}
$$

And hence its positive scalar $-b/\gamma \in N_{C_1}(x)$ as well.

Finally, using the fact $(x, 0) \in E_1$, and by definition $(\forall y \in C_2)\ (y, \langle n, y-x\rangle) \in E_2$. Substitute those into $(\mathrm{II.})$ again we get

$$
\begin{aligned}
&(\forall y \in C_2)\ \langle b, x\rangle \leq \langle b, y\rangle + \gamma\langle n, y-x\rangle \\
\iff &(\forall y \in C_2)\ \left\langle \frac{b}{\gamma} + n, y-x \right\rangle \leq 0 \\
\iff &\frac{b}{\gamma} + n \in N_{C_2}(x)
\end{aligned}
$$

As such $n \in N_{C_1}(x) + N_{C_2}(x)$, and we're done. □

Proposition L7-d: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Let $x \in \mathbb{R}^m$ and let $u \in \mathbb{R}^m$. Then

$$
u \in \partial f(x) \iff (u, -1) \in N_{\mathrm{epi}(f)}(x, f(x))
$$

Proof: Observe that $\mathbf{epi}(f) \neq \emptyset$ and is convex. (because f is proper and convex). Now let $u \in \mathbb{R}^m$. Then

$$
\begin{aligned}
&(u, -1) \in N_{\mathrm{epi}(f)}(x, f(x)) \\
\iff &x \in \mathrm{dom}(f) \text{ and } (\forall (y, \beta) \in \mathrm{epi}(f))\ \langle (y-x, \beta - f(x)), (u, -1)\rangle \leq 0 \\
\iff &x \in \mathrm{dom}(f) \text{ and } (\forall (y, \beta) \in \mathrm{epi}(f))\ \langle y-x, u\rangle + f(x) \leq \beta \\
\iff &x \in \mathrm{dom}(f) \text{ and } (\forall y \in \mathrm{dom}(f))\ \langle y-x, u\rangle + f(x) \leq f(y) \\
\iff &u \in \partial f(x)
\end{aligned}
$$

This is easily visualizable as well. □

Theorem 23.8, the big one: Let $f : \mathbb{R}^m \to (-\infty, +\infty], g : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Suppose taht $\mathbf{ri}(\mathrm{dom}(f)) \cap \mathbf{ri}(\mathrm{dom}(g)) \neq \emptyset$. Then $(\forall x \in \mathbb{R}^m)$ we have

$$
\partial f(x) + \partial g(x) = \partial(f+g)(x)
$$

Proof: Let $x \in \mathbb{R}^m$. If $x \notin \mathrm{dom}(f) \cap \mathrm{dom}(g) = \mathrm{dom}(f+g) \supseteq \mathrm{dom}(\partial f) \cap \mathrm{dom}(\partial g)$. Then $\partial f(x) + \partial g(x) = \emptyset$. Also, $\partial(f+g)(x) = \emptyset$.

Now, let $x \in \mathrm{dom}(f) \cap \mathrm{dom}(g) = \mathrm{dom}(f+g)(x)$. One can easily verify that

$$
\partial f(x) + \partial g(x) \subseteq \partial(f+g)(x)
$$

We now verify the opposite inclusion. Suppose that $u \in \partial(f+g)(x)$. Then

$$
(\forall y \in \mathbb{R}^m)\ (f+g)(y) \geq (f+g)(x) + \langle u, y-x\rangle \qquad (\mathrm{I.})
$$

Consider 2 nonempty closed convex sets,

$$
\begin{cases}
E_1 = \{(x, \alpha, \beta) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} : f(x) \leq \alpha\} = \mathrm{epi}(f) \times \mathbb{R} \\
E_2 = \{(x, \alpha, \beta) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} : g(x) \leq \beta\}
\end{cases}
$$

We claim that

$$
(u, -1, -1) \in N_{E_1 \cap E_2}(x, f(x), g(x)) \qquad (\mathrm{II.})
$$

Indeed, let $(y, \alpha, \beta) \in E_1 \cap E_2$, then

$$
\begin{aligned}
&\langle (u, -1, -1),\ (y, \alpha, \beta) - (x, f(x), g(x))\rangle \\
=\ &\langle u, y-x\rangle - (\alpha - f(x)) - (\beta - g(x)) \\
=\ &\langle u, y-x\rangle + (f+g)(x) - (\alpha + \beta) \\
\leq\ &(f+g)(y) - \alpha - \beta \qquad \text{by } (\mathrm{I.}) \\
=\ &f(y) - \alpha + g(y) - \beta \leq 0
\end{aligned}
$$

Which proves $(\mathrm{II.})$ by definition of normal cone.

Next we claim that $\mathbf{ri}(E_1) \cap \mathbf{ri}(E_2) \neq \emptyset$ using the earlier result, Fact L7-b, we get

$$\begin{cases} \text{ri}(E_1) = \text{ri}(\text{epi}(f)) \times \text{ri}(\mathbb{R}) = \text{ri}(\text{epi}(f)) \times \mathbb{R} \\ \text{ri}(E_2) = \{(x, \alpha, \beta) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{R} : x \in \text{ri}(\text{dom}(g)), g(x) < \beta\} \end{cases}$$

Let $z \in \text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(g))$. Then $(z, f(z)+1, g(z)+1) \in \text{ri}(E_1) \cap \text{ri}(E_2) \neq \emptyset$. Therefore, E_1, E_2 are nonempty, closed convex with $\text{ri}(E_1) \cap \text{ri}(E_2) \neq \emptyset$. So by an earlier result Theorem L7-c, we have

$$N_{E_1 \cap E_2}(x, f(x), g(x)) = N_{E_1}(x, f(x), g(x)) + N_{E_2}(x, f(x), g(x))$$

Therefore, we must have (see special note):

$$\begin{aligned} (u, -1, -1) = & \underbrace{(u_1, -1, 0)}_{\in N_{E_1}(x, f(x), g(x))} + \underbrace{(u_2, 0, -1)}_{\in N_{E_2}(x, f(x), g(x))} \\ u = & \, u_1 + u_2 \end{aligned}$$

> Take special note that $N_{E_1}(x, \alpha, \beta) = N_{\text{epi}(f)}(x, \alpha) \times N_{\mathbb{R}}(\beta) = N_{\text{epi}(f)}(x, \alpha) \times \{0\}$. This is left as assignment exercise. Also recall that when a point is in an interior of the set, the normal cone at that point is always $\{0\}$.

$$\begin{cases} (u_1, -1) \in N_{\text{epi}(f)}((x, f(x))) \\ (u_2, -1) \in N_{\text{epi}(g)}((x, g(x))) \end{cases}$$

Recalling an earlier result, Proposition L7-d, we conclude that

$$\begin{cases} u_1 \in \partial f(x) \\ u_2 \in \partial g(x) \end{cases} \implies u = u_1 + u_2 \in \partial f(x) + \partial g(x)$$

Proof is now complete. □

Important Example L7-e: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. adn proper, and let $\emptyset \neq C \subseteq \mathbb{R}^m$ be convex and closed. Suppose that $\text{ri}(C) \cap \text{ri}(\text{dom}(f)) \neq \emptyset$.

Consider the problem:

$$(P) \qquad \min(f(x) : x \in C)$$

Let $\bar{x} \in \mathbb{R}^m$. Then $\bar{x}$ solves (P) if and only if $(\partial f(\bar{x})) \cap (-N_C(\bar{x})) \neq \emptyset$.

> Take some time to visualize this geometrically. It's normal for $\partial f(\bar{x})$ (direction of steepest ascend) to be in the normal cone. However in this case, note that the normal cone is negative.

Proof: write (P) as $\min(f(x) + \delta_C(x) : x \in \mathbb{R}^m)$. Observe that $f + \delta_C$ is convex l.s.c. and proper.

By Fermat's theorem, $\bar{x}$ solves P if and only if $0 \in \partial(f + \delta_C)(\bar{x})$. Now,

$$\text{ri}(\text{dom}(f)) \cap \text{ri}(\text{dom}(\delta_C)) = \text{ri}(\text{dom}(f)) \cap \text{ri}(C) \neq \emptyset$$

Therefore, by an earlier Theorem 23.8, we conclude that

$$\begin{aligned} \bar{x} \text{ solves } P &\iff 0 \in \partial(f + \delta_C)(\bar{x}) = \partial f(\bar{x}) + \partial \delta_C(\bar{x}) = \partial f(\bar{x}) + N_C(\bar{x}) \\ &\iff (\exists u \in \partial f(\bar{x})) \text{ such that } (-u) \in N_C(\bar{x}) \\ &\iff \partial f(\bar{x}) \cap (-N_C(\bar{x})) \neq \emptyset \end{aligned}$$

And we're done. □

Important Example L7-f: Let $d \in \mathbb{R}^m$, and let $\emptyset \neq C \subseteq \mathbb{R}^m$ be convex and closed. Consider the problem

$$(P) \qquad \min(\langle d, x \rangle : x \in C)$$

Let $\bar{x} \in \mathbb{R}^m$. Then $\bar{x}$ solves P if and only if $-d \in N_C(\bar{x})$.

Differentiability of Convex Function

Directional Derivative: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper, and let $x \in \text{dom}(f)$. The directional derivative of f at x in the direction of d is

$$f'(x, d) := \lim_{t \to 0} \frac{f(x + td) - f(x)}{t}$$

Gradient Operator: f is **differentiable** at x if there exists a linear operator $\nabla f(x) : \mathbb{R}^m \to \mathbb{R}^m$ called the gradient of f at x that satisfies: for any $y \neq 0$

$$\lim_{\|y\| \to 0} \frac{\|f(x+y) - f(x) - \nabla f(x) \cdot y\|}{\|y\|} = 0$$

> Remark: $f'(x, d) = \langle \nabla f(x), d \rangle$

Theorem 23.2: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex and proper, and let $x \in \mathrm{dom}(f)$. Then u is a subgradient of f if and only if

$$(\forall y \in \mathbb{R}^m) \quad f'(x, y) \geq \langle u, y \rangle$$

Proof: Using the subgradient inequality we have

$$\begin{aligned} u \in \partial f(x) &\iff (\forall y \in \mathbb{R}^m)(\forall \lambda > 0)\ f(x + \lambda y) \geq f(x) + \langle u, x + \lambda y - x \rangle \\ &\iff (\forall y \in \mathbb{R}^m)(\forall \lambda > 0)\ \frac{f(x + \lambda y) - f(x)}{\lambda} \geq \langle u, y \rangle \end{aligned}$$

Taking the limit as $\lambda \to 0$ yields the desired result. □

Theorem 25.2: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex and proper and let $x \in \mathrm{dom}(f)$. If f is differentiable at x, then $\nabla f(x)$ is the unique subgradient of f at x.

Proof: Recall that $(\forall y \in \mathbb{R}^m)\ f'(x, y) = \langle \nabla f(x), y \rangle$. Let $u \in \mathbb{R}^m$. Using Theorem 23.2, we have $u \in \partial f(x) \iff (\forall y \in \mathbb{R}^m)\ f'(x, y) \geq \langle u, y \rangle \iff (\forall y \in \mathbb{R}^m) \langle \nabla f(x), y \rangle \geq \langle u, y \rangle$. Clearly, we have $\{\nabla f(x)\} \subseteq \partial f(x)$.

Moreover, letting $y = u - \nabla f(x)$ yields $\|u - \nabla f(x)\|^2 \leq 0 \implies u = \nabla f(x) \implies \partial f(x) \subseteq \{\nabla f(x)\}$ (trivially). And hence $\partial f(x) = \{\nabla f(x)\}$. □

Lemma L8-a: Let $\phi : \mathbb{R} \to (-\infty, +\infty]$ be a proper function that is differentiable on a non-empty open interval $I \subseteq \mathrm{dom}(\phi)$. Then: ϕ' is **increasing** on I implies ϕ is convex on I.

Proof: fixate $x \in I$ and $\lambda \in (0, 1)$. Set $\psi : \mathbb{R} \to (-\infty, +\infty] : z \mapsto \lambda\phi(x) + (1-\lambda)\phi(z) - \phi(\lambda x + (1-\lambda)z)$.

Then $\psi'(z) = (1-\lambda)\phi'(z) - (1-\lambda)\phi'(\lambda x + (1-\lambda)z)$. Because ϕ' is increasing, we get

$$\begin{cases} \psi'(x) = 0 = \psi(x) & \\ \psi'(z) < 0, & z < x \\ \psi'(z) > 0, & z > x \end{cases}$$

Therefore using first **derivative test**, ψ achives its infimum on I at x. That is, $(\forall y \in I)\ \psi(y) \geq \psi(x) = 0$. Finally, we arrive at the convex characterization inequality: $(\forall y \in I)\ \lambda\phi(x) + (1-\lambda)\phi(y) \geq \phi(\lambda x + (1-\lambda)y)$. □

Proposition L8-b, Function Convexity Characterization using Gradient: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper. Suppose that $\mathrm{dom}(f)$ is open and convex, and that f is differentiable on $\mathrm{dom}(f)$. Then the following are equivalent:

1. f is convex.
2. Gradient satisfies subgradient inequality:

$$(\forall x \in \mathrm{dom}(f))(\forall y \in \mathrm{dom}(f))\ \langle x - y, \nabla f(y) \rangle + f(y) \leq f(x)$$

3. Gradient is monotone:

$$(\forall x \in \mathrm{dom}(f))(\forall y \in \mathrm{dom}(f))\ \langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq 0$$

Proof:

- $(1. \implies 2.)$: follows directly from Theorem 25.2 and the subgradient inequality.
- $(2. \implies 3.)$: left as exercise (assignment).
- $(3. \implies 1.)$: see below, it's big.

Fix $x, y \in \mathrm{dom}(f), z \in \mathbb{R}^m$. By assumption, $\mathrm{dom}(f)$ is open. Therefore, $(\exists \epsilon > 0)$ such that $y + \epsilon(x - y) \in \mathrm{dom}(f)$ and $y + \epsilon(y - x) \in \mathrm{dom}(f)$. Set $C = (-\epsilon, 1 + \epsilon) \subseteq \mathbb{R}$, by convexity $(\forall \alpha \in C)x + \alpha(x - y) \in \mathrm{dom}(f)$. Now set $\phi = \mathbb{R} \to (-\infty, +\infty]$ defined by

$$\phi(\alpha) = f(y + \alpha(x - y)) + \delta_C(\alpha) \tag{I.}$$

The goal is to show that ϕ' is increasing on C. Because ϕ is differentiable on C,

$$\phi'(\alpha) = \langle \nabla f(y + \alpha(x-y)), x - y \rangle \tag{II.}$$

Now, take $\alpha, \beta \in C$ with $\alpha < \beta$. Set

$$\begin{cases} y_\alpha = y + \alpha(x-y) \\ y_\beta = y + \beta(x-y) \end{cases} \implies y_\beta - y_\alpha = (\beta - \alpha)(x-y)$$

Then

$$\begin{aligned} \phi'(\beta) - \phi'(\alpha) &= \langle \nabla f(y + \beta(x-y)), x-y \rangle - \langle \nabla f(y + \alpha(x-y)), x-y \rangle \\ &= \langle \nabla(y_\beta) - \nabla f(y_\alpha), x - y \rangle \\ &= \left\langle \nabla(y_\beta) - \nabla f(y_\alpha), \frac{y_\beta - y_\alpha}{\beta - \alpha} \right\rangle \\ &= \frac{1}{\beta - \alpha} \underbrace{\langle \nabla f(y_\beta) - \nabla f(y_\alpha), y_\beta - y_\alpha \rangle}_{\geq 0, \text{ given monotonicity}} \geq 0 \end{aligned}$$

That is, ϕ' is increasing on C, and by Lemma L8-a, ϕ is convex on C. Consequently, by definition of ϕ, we have

$$\begin{aligned} f(\alpha x + (1-\alpha)y) &= \phi(\alpha) \\ &\leq \alpha\phi(1) + (1-\alpha)\phi(0) \\ &= \alpha f(x) + (1-\alpha) f(y) \end{aligned}$$

which is the convex characterization inequality. □

Interestin Example L8-5: Let A be $m \times m$ matrix, and set $f : \mathbb{R}^m \to \mathbb{R}$, $f(x) = \langle x, Ax \rangle$. Then the following hold:

1. $\nabla f(x) = (A + A^\intercal)x$
2. f is convex if and only if $A + A^\intercal$ is positive semidefinite.

Proof for Part 2: By proposition L8-b part 1 and 3, we have

$$\begin{aligned} f \text{ is convex} &\iff (\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0 \\ &\iff (\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \langle (A + A^\intercal)x - (A + A^\intercal)y, x - y \rangle \geq 0 \\ &\iff (\forall z \in \mathbb{R}^m) \langle (A + A^\intercal)z, z \rangle \geq 0 \end{aligned}$$

And we're done. □

Subdifferentiability and Conjugacy

Proposition L8-c: Let f, g be functions from $\mathbb{R}^m$ to $[-\infty, +\infty]$. Then

1. The **bi-conjugate** $f^{**} := (f^*)^* \leq f$
2. $f \leq g \implies f^* \geq g^*$ and $f^{**} \leq g^{**}$.

Proposition L8-d: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper. Then $(\forall x \in \mathbb{R}^m)(\forall u \in \mathbb{R}^m)$, we have the **Frenchel-Young Inequality**:

$$f(x) + f^*(u) \geq \langle x, u \rangle$$

Proof: Observe that the definition of f^* yields:

$$f \equiv +\infty \iff f^* \equiv -\infty$$

However by assumption, proper function must have non-empty domain, and therefore $\exists \bar{x} \in \mathrm{dom}(f)$ such that $f(\bar{x}) < +\infty$, and therefore $(\forall u \in \mathbb{R}^m)\ f^*(u) \neq -\infty$. Therefore for $x \in \mathbb{R}^m$ such that $f(x) = +\infty$, the desired inequality clearly holds.

Else if $f(x) < +\infty$, the result follows from definition of convex conjugacy.

Proposition L8-d: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Let $x \in \mathbb{R}^m$ and let $u \in \mathbb{R}^m$. Then the following are equivalent:

$$u \in \partial f(x) \iff f(x) + f^*(u) = \langle x, u \rangle$$

Proof:

$$
\begin{aligned}
u \in \partial f(x) &\iff (\forall y \in \mathrm{dom}(f))\ \langle y - x, u \rangle + f(x) \leq f(y) \\
&\iff (\forall y \in \mathrm{dom}(f))\ \langle y, u \rangle - f(y) \leq \langle x, u \rangle - f(x) \leq f^*(u) \\
&\iff f^*(u) = \sup_{y \in \mathbb{R}^m} (\langle y, u \rangle - f(y)) \leq \langle x, u \rangle - f(x) \leq f^*(u) \\
&\iff f(x) + f^*(u) = \langle x, u \rangle
\end{aligned}
$$

And we are done. □

Proposition L8-e: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex and proper, let $x \in \mathbb{R}^m$ and suppose that $\partial f(x) \neq \emptyset$. Then

$$
f^{**}(x) = f(x)
$$

where $f^{**}(x) = \sup_{y \in \mathbb{R}^m} \{\langle y, x \rangle - f^*(y)\}$.

> $f^{**}(x) = f(x)$ holds for all convex, l.s.c., proper functions (see fact L8-f). However, the proof is much more complicated and is skipped here.

Proof: Let $u \in \partial f(x)$. By proposition L8-d,

$$
f(x) = \langle u, x \rangle - f^*(u)
$$

Consequently,

$$
\begin{aligned}
f^{**}(x) &= \sup_{y \in \mathbb{R}^m} \{\langle x, y \rangle - f^*(y)\} \\
&\geq \langle x, u \rangle - f^*(u) = f(x)
\end{aligned}
$$

Conversely,

$$
\begin{aligned}
f^{**}(x) &= \sup_{y \in \mathbb{R}^m} \{\langle x, y \rangle - f^*(y)\} \\
&= \sup_{y \in \mathbb{R}^m} \left\{ \langle x, y \rangle - \sup_{z \in \mathbb{R}^m} \{\langle z, y \rangle - f(z)\} \right\} \\
&= \sup_{y \in \mathbb{R}^m} \left\{ \langle x, y \rangle + \inf_{z \in \mathbb{R}^m} \{f(z) - \langle z, y \rangle\} \right\} \\
&= \sup_{y \in \mathbb{R}^m} \left\{ \inf_{z \in \mathbb{R}^m} \{f(z) + \langle x - z, y \rangle\} \right\} \\
&\leq \sup_{y \in \mathbb{R}^m} \{f(x) + \langle y, x - x \rangle\} \\
&= \sup_{y \in \mathbb{R}^m} f(x) = f(x)
\end{aligned}
$$

Altogether, $f(x) = f^{**}(x)$. □

Fact L8-f: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper. Then f is convex and l.s.c. if and only if

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

In this case, f^* is also proper.

Corollary L8-g: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Then

1. f^* is convex, l.s.c. and proper
2. $f^{**} = f$

Proposition L8-h: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Then

$$
u \in \partial f(x) \iff x \in \partial f^*(u)
$$

Proof: Set $g = f^*$, by propostion L8-d and a serious of propositions

$$
\begin{aligned}
u \in \partial f(x) &\iff f(x) + f^*(u) = \langle x, u \rangle \\
&\iff g^*(x) + g(u) = \langle x, u \rangle \\
&\iff x \in \partial g(u) = \partial f^*(u)
\end{aligned}
$$

And we're done. □

Theorem L9-a: Let $f : \mathbb{R}^m \to \mathbb{R}$ be proper, l.s.c. and let C be a compact subset of $\mathbb{R}^m$ such that $C \cap \mathrm{dom}(f) \neq \emptyset$. Then the following hold:

1. f is bounded below over C
2. f attains its minimal value over C

Proof: 1. suppose for eventual contradiction that f is not bounded below over C. Then there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in C such that $\lim_{n\to\infty} f(x_n) = -\infty \quad (\mathrm{I.})$. Recall that C is compact, equivalently, C is closed and bounded (finite dimension). Hence $(x_n)_{n\in\mathbb{N}}$ must be bounded. By bolzano-Weierstrass Theorem, there exists a convergent subsequence, say $x_{k_n} \to \bar{x} \in C$.

Since f is l.s.c., we learn that

$$f(\bar{x}) \leq \liminf_{n\to\infty} f(x_{k_n})$$

which is absurd in view of $(\mathrm{I.})$.

For point 2, let $f_{\min}$ be the minimal(infimum) value of f over C. Then there exists a sequence $(x_n)_{n\in\mathbb{N}}$ in C such that

$$f(x_n) \to f_{\min}$$

Since C is bounded, $(x_n)_{n\in\mathbb{N}}$ is also bounded. Let $\bar{x}$ be a **cluster point** of $(x_n)_{n\in\mathbb{N}}$, say $x_{k_n} \to \bar{x} \in C$ by closure. Then by l.s.c.

$$f(\bar{x}) \leq \liminf_{n\to\infty} f(x_{k_n}) = f_{\min}$$

Hence, $\bar{x}$ is a minimizer of f over C. □

Coercive Function: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$. Then f is coercive if

$$\lim_{\|x\|\to+\infty} f(x) = +\infty$$

and f is **super coercive** if

$$\lim_{\|x\|\to+\infty} \frac{f(x)}{\|x\|} = +\infty$$

Theorem L9-b: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper, l.s.c. and coercive and let C be a closed subset of $\mathbb{R}^m$ satisfying that $C \cap \mathrm{dom}(f) \neq \emptyset$. Then f attains its minimal value over C.

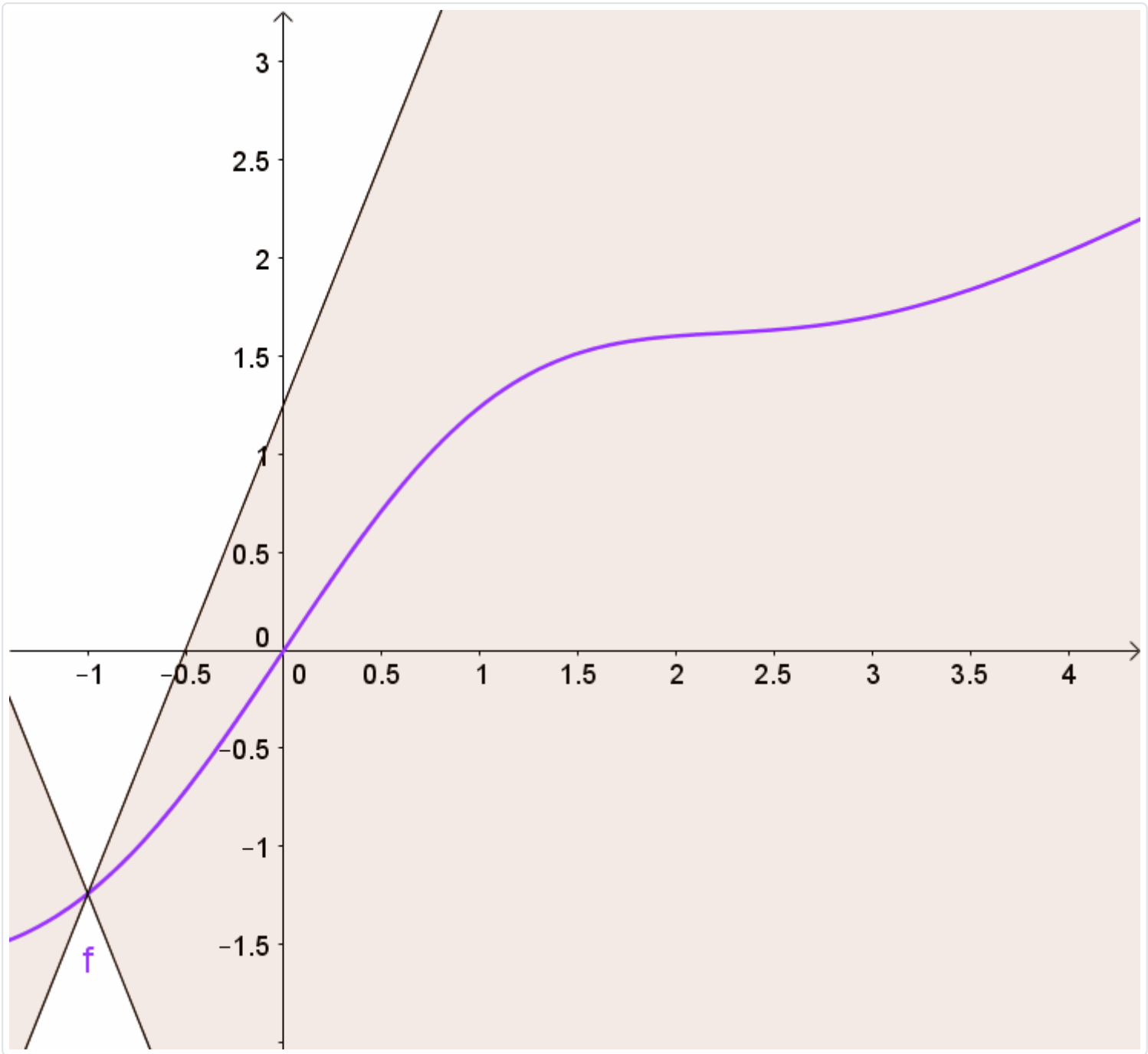
Proof: Let $x \in C \cap \mathrm{dom}(f)$. Since f is coercive, $(\exists M > 0)$ such that $f(y) > f(x)$ whenever $\|y\| > M$. Observe that if $\bar{f}$ is a minimizer of f over C, we have $f(\bar{x}) \leq f(x)$. We then know that the set of minimizers of f over C is the same as the set of minimizers of f over $C \cap B(0, M)$. The latter is closed and bounded hence compact. We simply apply Theorem L9-a to arrive at the result. □

Differentiability and Strong Convexity

L-Lipschitz Continuity: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ and let $L \geq 0$. Then T is L-Lipschitz if $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$

$$\|T(x) - T(y)\| \leq L\|x - y\|$$

To visualize this type of continuity, we quote wikipedia: For a Lipschitz continuous function, there exists a double cone (white) whose origin can be moved along the graph so that the whole graph always stays outside the double cone.



Example L9-5: Let $f : \mathbb{R}^m \to \mathbb{R}^m : x \mapsto 0.5\langle x, Ax\rangle + \langle b, x\rangle + c$, where A is positive semidefinite, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}$. Then the following hold:

1. $(\forall x \in \mathbb{R}^m)\ \nabla f(x) = Ax + b$
2. ∇f is Lipschitz continuous with a constant $L = \|A\|$ where $\|A\| = \max\limits_{\|x\|\neq 0} \dfrac{\|Ax\|}{\|x\|}$

Proof: Part 1 follows from L8-5 that $(\forall x \in \mathbb{R}^m)\nabla f(x) = 0.5(A + A^\intercal)x + b = 0.5(A + A)x + b = Ax + b$.

For the second part, indeed, $\|\nabla f(x) - \nabla f(y)\| = \|Ax - Ay\| = \|A(x - y)\| \leq \|A\|\ \|x - y\|$. And the conclusion follows. □

Example L9-6: Let C be a nonempty closed convex subset of $\mathbb{R}^m$. Then P_C is Lipschitz continuous at constant 1.

Proof: If C is a singleton, the conclusion is trivial. Now suppose that C is not a singleton. Let $\{x, y\} \subseteq \mathbb{R}^m$ such that $x \neq y$. If $P_C(x) = P_C(y)$ then $0 = \|P_C(x) - P_C(y)\| < \|x - y\|$. Else if $P_C(x) \neq P_C(y)$, then:

$$
\begin{aligned}
&\|P_C(x) - P_C(y)\|^2 \\
=&\langle P_C(x) - P_C(y), P_C(x) - P_C(y)\rangle \\
=&\underbrace{\langle P_C(x) - P_C(y), P_C(x) - x\rangle}_{\leq 0,\ \text{projection theorem}} + \underbrace{\langle P_C(x) - P_C(y), y - P_C(y)\rangle}_{\leq 0,\ \text{projection theorem}} + \langle P_C(x) - P_C(y), x - y\rangle \\
\leq&\langle P_C(x) - P_C(y), x - y\rangle \\
\leq&\|P_C(x) - P_C(y)\|\ \|x - y\| \qquad \text{by Cauchy-Schwarz}
\end{aligned}
$$

Dividing both sides of $\|P_C(x) - P_C(y)\|$ (which is non-zero), and we get $\|P_C(x) - P_C(y)\| \leq \|x - y\|$. □

Descent Lemma, L9-c: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be differentiable on $\emptyset \neq D \subseteq \operatorname{int}(\operatorname{dom}(f))$ such that ∇f is L-Lipschitz over D and D is convex. Then $(\forall x \in D)(\forall y \in D)$ we have

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

Proof: Recall that the fundamental theorem of calculus implies that

$$
\begin{aligned}
f(y) - f(x) &= \int_0^1 \langle \nabla f(x + t(y-x)), y - x \rangle \, dt \\
&= \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y-x)) - \nabla f(x), y - x \rangle \, dt
\end{aligned}
$$

Hence,

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

Hence we arrive at the result. □

Theorem L9-d, characterization of L-Lipschitz continuous gradient: Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and differentiable, and let $L > 0$. Then the following are equivalent:

1. ∇f is L-Lipschitz.
2. $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|^2$
3. $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2$
4. $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2$

Proof: $(1. \implies 2.)$: follows from descent lemma.

$(2. \implies 3.)$: We assume that $\nabla f(x) \neq \nabla f(y)$. Otherwise, the conclusion follows immediately using the subgradient inequality.

Fix $x \in \mathbb{R}^m$ and set: $h_x : \mathbb{R}^m \to \mathbb{R}$ with $h_x(y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$. Observe that h_x is convex (sum of convex functions) and differentiable, and $\nabla h_x(y) = \nabla f(y) - \nabla f(x)$.

We claim that $(\forall y \in \mathbb{R}^m)(\forall z \in \mathbb{R}^m)\ h_x(z) \leq h_x(y) + \langle \nabla h_x(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \quad (\mathbf{I.})$. Indeed, this follows from point 2., the definition of h_x, and some clever manipulations.

Observe that $\nabla h_x(x) = \nabla f(x) - \nabla f(x) = 0$. Hence, because h_x is convex, x is a global minimizer of h_x, by Fermat's Theorem.

Let $y \in \mathbb{R}^m$ and let $v \in \mathbb{R}^m$ be such that $\|v\| = 1$ and $\langle \nabla h_x(y), v \rangle = \|\nabla h_x(y)\|$. Set $z = y - \frac{\|\nabla h_x(y)\|}{L} v$, and since x is minimizer of h_x, we have

$$
\begin{aligned}
0 = h_x(x) &\leq h_x \left(y - \frac{\|\nabla h_x(y)\|}{L} v \right) \\
&\leq h_x(y) - \frac{\|\nabla h_x(y)\|}{L} \langle \nabla h_x(y), v \rangle + \frac{1}{2L} \|\nabla h_x(y)\|^2 \, \|v\|^2 \\
&= h_x(y) - \frac{\|\nabla h_x(y)\|^2}{L} + \frac{1}{2L} \|\nabla h_x(y)\|^2 \\
&= h_x(y) - \frac{1}{2L} \|\nabla h_x(y)\|^2 \\
&= f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{1}{2L} \|\nabla h_x(y)\|^2
\end{aligned}
$$

$(3. \implies 4.)$: using point 3 we have

$$
\begin{cases}
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 \\
f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2
\end{cases}
$$

Adding the above two inequalities yield point 4.

$(4. \implies 1.)$: We assume that $\nabla f(x) \neq \nabla f(y)$ (otherwise the conclusion is trivial). Now point 4. implies

$$\begin{aligned}\|\nabla f(x)-\nabla f(y)\|^2 &\le L\langle \nabla f(x)-\nabla f(y), x-y\rangle \\ &\le L\|\nabla f(x)-\nabla f(y)\|\,\|x-y\|\end{aligned}$$

This implies that $\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|$ and hence the Lipschitz continuous gradient. □

Example L10-1, **firm nonexpansiveness**: Let C be nonempty closed convex subset of $\mathbb{R}^m$, then $(\forall x\in\mathbb{R}^m)(\forall y\in\mathbb{R}^m)\ \|P_C(x)-P_C(y)\|^2 \le \langle P_C(x)-P_C(y), x-y\rangle$

Proof Hint: rearrange and use projection theorem. □

Example L10-2: Let C be nonempty closed and convex subset of $\mathbb{R}^m$. Consider the function $f:\mathbb{R}^m\to\mathbb{R}$ such that $f(x)=0.5d_C^2(x)$, then the following holds

1. f is differentiable over $\mathbb{R}^m$ and $(\forall x\in\mathbb{R}^m)\ \nabla f(x)=x-P_C(x)$
2. ∇f is 1-Lipschitz

Proof: 1. Let $x\in\mathbb{R}^m$. Define $(\forall y\in\mathbb{R}^m)\ h_x(y)=f(x+y)-f(x)-\langle y, x-P_C(x)\rangle$. Clearly, h_x is convex.

By definition of $\nabla f(x)$ (that $\lim_{h\to 0}\frac{|f(x+h)-f(x)-\nabla(x)\cdot h|}{\|h\|}=0$), it is sufficient to show that $\frac{|h_x(y)|}{\|y\|}\to 0$ as $y\to 0$.

Observe that $(\forall x\in\mathbb{R}^m)\ f(x)=\frac{1}{2}d_C^2(x)=\frac{1}{2}\|x-P_C(x)\|^2$. Now on the one hand,

$$\begin{aligned}h_x(y)&=\frac{1}{2}\|(x+y)-P_C(x+y)\|^2-\frac{1}{2}\|x-P_C(x)\|^2-\langle y, x-P_C(x)\rangle \\ &\le \frac{1}{2}\|(x+y)-P_C(x)\|^2-\frac{1}{2}\|x-P_C(x)\|^2-\langle y, x-P_C(x)\rangle \\ &=\left(\frac{1}{2}\|x-P_C(x)\|^2+\langle y, x-P_C(x)\rangle+\frac{1}{2}\|y\|^2\right)-\frac{1}{2}\|x-P_C(x)\|^2-\langle y, x-P_C(x)\rangle \\ &=\frac{1}{2}\|y\|^2 \qquad (\mathrm{I.})\end{aligned}$$

On the other hand, by the above argument $h_x(-y)\le\frac{1}{2}\|y\|^2$. Therefore, $0=h_x(0)=h_x(\frac{1}{2}(y+(-y)))\le\frac{1}{2}h_x(y)+\frac{1}{2}h_x(-y)$ by convexity, and this implies $h_x(y)\ge -h_x(-y)\ge-\frac{1}{2}\|y\|^2 \quad (\mathrm{II.})$. Consequently, $|h_x(y)|\le\frac{1}{2}\|y\|^2$. And hence $\frac{|h_x(y)|}{\|y\|}\to 0$ as $y\to 0$ by the squeeze theorem.

For Part 2, let $x\in\mathbb{R}^m, y\in\mathbb{R}^m$.

$$\begin{aligned}\|\nabla f(x)-\nabla f(y)\|^2 &= \|x-P_C(x)-(y-P_C(y))\|^2 \\ &= \|x-y\|^2-2\langle x-y, P_C(x)-P_C(y)\rangle+\|P_C(x)-P_C(y)\|^2 \\ &\le \|x-y\|^2-2\|P_C(x)-P_C(y)\|^2+\|P_C(x)-P_C(y)\|^2 \quad (\text{L10-1}) \\ &= \|x-y\|^2-\|P_C(x)-P_C(y)\|^2 \\ &\le \|x-y\|^2\end{aligned}$$

Hence the 1-Lipschitz continuous gradient. □

Theorem L10-a, Second Order Characterization of Lipschitz Continuous Gradient: Let $f:\mathbb{R}^m\to\mathbb{R}$ be twice continuously differentiable over $\mathbb{R}^m$, and let $L\ge 0$. Then the following are equivalent:

1. ∇f is L-Lipschitz.
2. $(\forall x\in\mathbb{R}^m)\ \|\nabla^2 f(x)\|\le L$, note that $\nabla^2 f$ is the **Hessian matrix** of f.

Recall Hessian matrix :'(

$$\mathbf{H}f = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix},$$

Warning: the proof is super hard to comprehend because of calculus.

Proof: $(1. \implies 2.)$: Suppose that ∇f is L-Lipschitz continuous. Observe that for any $y \in \mathbb{R}^m, \alpha > 0$. We have $\|\nabla f(x+\alpha y) - \nabla f(x)\| \leq \alpha L \|y\|$. That is, the second-order directional derivative (makes sense if you break down the Hessian row-by-row, and assemble the final vector, it's basically the norm of vector of directional derivatives for $\partial f / \partial x_i$)

$$\begin{aligned} \|[\nabla^2 f(x)]\, y\| &= \lim_{\alpha \to 0^+} \frac{\|\nabla f(x+\alpha y) - \nabla f(x)\|}{\alpha} \\ &\leq \lim_{\alpha \to 0^+} \frac{\alpha L \|y\|}{\alpha} = L\|y\| \end{aligned}$$

Equivalently, $\|\nabla^2 f(x)\| \leq L$ as desired.

$(2. \implies 1.)$: Suppose that for any $x \in \mathbb{R}^m$, $\|\nabla^2 f(x)\| \leq L$. Using the fundamental theorem of calculus we have $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$

$$\begin{aligned} \nabla f(x) &= \nabla f(y) + \int_0^1 \left[\nabla^2 f(y + \alpha(x-y))\right]\ (x-y) d\alpha \\ &= \nabla f(y) + \left[\int_0^1 \nabla^2 f(y+\alpha(x-y))\, d\alpha\right]\ (x-y) \\ \|\nabla f(x) - \nabla f(y)\| &= \left\| \left[\int_0^1 \nabla^2 f(y+\alpha(x-y))\, d\alpha\right] (x-y) \right\| \\ &\leq \left\| \left[\int_0^1 \nabla^2 f(y+\alpha(x-y))\, d\alpha\right] \right\|\ \|x-y\| \quad \text{easy to prove} \\ &\leq \left(\int_0^1 \left\| \left[\nabla^2 f(y+\alpha(x-y))\ \right] \right\| d\alpha \right)\ \|x-y\| \\ &\leq L\|x-y\| \qquad \text{by 2.} \end{aligned}$$

which proves L-Lipschitz continuous gradient. □

Fact L10-4: Let A be a $m \times m$ symmetric matrix. Then

$$\sup_{\|x\|=1} \|Ax\| = \max(\{\lambda_1, \ldots, \lambda_m\})$$

where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of A.

Proposition L10-b: Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and twice continuously differentiable. Then f is convex if and only if $(\forall x \in \mathbb{R}^m)\ \nabla^2 f(x)$ is positive semi-definite.

Corollary L10-c: Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex and twice continuously differentiable and let $L \geq 0$. Then ∇f is L-Lipschitz if and only if $(\forall x \in \mathbb{R}^m)\ \lambda_{\max}(\nabla^2 f(x)) \leq L$ where $\lambda_{\max}$ means the maximum eigen value.

Proof: Since f is convex and twice continuously differentiable, we have $(\forall x \in \mathbb{R}^m)\ \nabla^2 f(x)$ is positive semidefinite. From Fact L10-4, we learn that

$$\begin{aligned} L &\geq \|\nabla^2 f(x)\| \\ &= |\lambda_{\max}(\nabla^2 f(x))| \qquad \text{linear algebra fact.} \\ &= \lambda_{\max}(\nabla^2 f(x)) \qquad \text{positive semidefinite} \end{aligned}$$

And we're done. □

> Example 10-7: Let $f : \mathbb{R}^m \to \mathbb{R}$ be given by $(\forall x \in \mathbb{R}^m)\ f(x) = \sqrt{1 + \|x\|^2}$. Prove that
>
> 1. f is convex
> 2. ∇f is 1-Lipschitz
>
> Proof Hint: For part 1, use Proposition L8-b, use Gradient Monotonicity to characterize convexity. For part 2. use Corollary L10-c for part 2.

Proposition L10-d, strong convexity characteriztion: Let $\beta > 0$. $f : \mathbb{R}^m \to (-\infty, +\infty]$ is β-strongly convex if and only if $f - \frac{\beta}{2}\|?\|^2$ is convex. (where $\|?\|$ is the norm function).

Proposition L10-e: Let $f : \mathbb{R}^m \to (-\infty, +\infty], g : \mathbb{R}^m \to (-\infty, +\infty]$, and let $\beta > 0$. Suppose that f is β-strongly convex and that g is convex. Then $(f + g)$ is β-strongly convex.

Proof: Set $h = f + g - \frac{\beta}{2}\|?\|^2 = \underbrace{(f - \frac{\beta}{2}\|?\|^2)}_{\text{convex by L10-d}} + g$. Then h is convex being the sum of two convex functions. We then apply L10-4 again, and arrive at the result. □

Fact L10-10: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Then f has a unique minimizer.

The Proximal Operator

Let $f : \mathbb{R}^m \to (-\infty, +\infty]$. The **proximal point mapping** of f is the operator

$$\begin{aligned} \operatorname{prox}_f : \mathbb{R}^m &\longrightarrow \mathcal{P}(\mathbb{R}^m) \\ x &\longmapsto \operatorname*{arg\,min}_{v \in \mathbb{R}^m} \left\{ f(v) + \frac{1}{2}\|v - x\|^2 \right\} \end{aligned}$$

Theorem L10-f: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Then $(\forall x \in \mathbb{R}^m)\ \operatorname{prox}_f(x)$ is a singleton.

Proof: Observe that for a fixed $x \in \mathbb{R}^m$, define $h_x = \frac{1}{2}\|? - x\|^2$, which is β-strongly convex for every $\beta < 1$. Let $g_x = f + h_x$, using Proposition L10-e, we know $(\forall x \in \mathbb{R}^m)\ g_x$ is strongly convex as well. Moreover, $(\forall x \in \mathbb{R}^m)\ g_x$ is l.s.c. and proper (using results from A2).

Therefore, apply Fact 10-10 we learn that $(\forall x \in \mathbb{R}^m)\ \operatorname*{arg\,min}_{v \in \mathbb{R}^m} g_x = \operatorname{prop}_f(x)$ exists and is unique. □

Example L10-13: Let C be a nonempty closed convex subset of $\mathbb{R}^m$. Then

$$\operatorname{prox}_{\delta_C} = P_C$$

Proof: Let $x \in \mathbb{R}^m$. By definition

$$\begin{aligned} & p \in \operatorname*{arg\,min}_{v \in \mathbb{R}^m} \left\{ \delta_C(v) + \frac{1}{2}\|x - v\|^2 \right\} \\ \iff & (\forall v \in \mathbb{R}^m)\ \delta_C(p) + \frac{1}{2}\|x - p\|^2 \leq \delta_C(v) + \frac{1}{2}\|x - u\|^2 \\ \iff & (\forall p \in C)(\forall v \in C)\ \|x - p\| \leq \|x - v\| \\ \iff & p = P_C(x) \end{aligned}$$

> Note that we may write $p \in \operatorname{prox}_f$ and $p = \operatorname{prox}_f$ interchangeably in the case of singleton.

Proposition L10-g, generalized characterization of projection: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Let $x \in \mathbb{R}^m, p \in \mathbb{R}^m$. Then $p = \operatorname{prox}_f(x)$ if and only if $(\forall y \in \mathbb{R}^m)\ \langle y - p, x - p \rangle + f(p) \leq f(y)$.

Proof: let $y \in \mathbb{R}^m$.

$(\Rightarrow)$: suppose that $p = \operatorname{prox}_f(x)$ and set $(\forall \lambda \in (0,1))\ p_\lambda = \lambda y + (1-\lambda)p$. Then,

$$\begin{aligned} f(p) + \frac{1}{2}\|x - p\|^2 &\leq f(p_\lambda) + \frac{1}{2}\|x - p_\lambda\|^2 \\ \implies f(p) &\leq f(p_\lambda) + \frac{1}{2}\|x - p_\lambda\|^2 - \frac{1}{2}\|x - p\|^2 \\ &= f(p_\lambda) + \frac{1}{2}\|x - \lambda y - (1-\lambda)p\|^2 - \frac{1}{2}\|x - p\|^2 \\ &= f(p_\lambda) + \frac{1}{2}\langle x - p - \lambda(y-p) - (x-p), x - p - \lambda(y-p) + (x-p) \rangle \\ &= f(p_\lambda) + \frac{\lambda^2}{2}\|y - p\|^2 - \lambda\langle x - p, y - p \rangle \\ &= f(\lambda y + (1-\lambda)p) + \frac{\lambda^2}{2}\|y - p\|^2 - \lambda\langle x - p, y - p \rangle \\ &\leq \lambda f(y) + (1-\lambda)f(p) + \frac{\lambda^2}{2}\|y - p\|^2 - \lambda\langle x - p, y - p \rangle \\ \implies \lambda\langle x - p, y - p \rangle + \lambda f(p) &\leq \lambda f(y) + \frac{\lambda^2}{2}\|y - p\|^2 \end{aligned}$$

Dividing both sides by λ and taking the limit $\lambda \to 0$ yields the desired inequality.

$(\Leftarrow)$: Suppose that $\langle y - p, x - p \rangle + f(p) \leq f(y)$, then

$$
\begin{aligned}
f(p) &\leq f(y) + \langle x-p, p-y\rangle \\
\implies f(p) + \frac{1}{2}\|x-p\|^2 &\leq f(y) + \langle x-p, p-y\rangle + \frac{1}{2}\|x-p\|^2 \\
&\leq f(y) + \langle x-p, p-y\rangle + \frac{1}{2}\|x-p\|^2 + \frac{1}{2}\|p-y\|^2 \\
&= f(y) + \frac{1}{2}\|(x-p)+(p-y)\|^2 \\
&= f(y) + \frac{1}{2}\|x-y\|^2
\end{aligned}
$$

And we're done. □

Example L10-15: Let $f : \mathbb{R} \to \mathbb{R} : x \mapsto |x|$. Then,

$$
\mathrm{prox}_f(x) = \begin{cases} x-1, & x > 1 \\ 0, & -1 \leq x \leq 1 \\ x+1, & x < -1 \end{cases}
$$

Proof: let $p \in \mathbb{R}^m$. Recall that $p = \mathrm{prox}_f(x)$ if and only if $(\forall y \in \mathbb{R})\ (y-p)(x-p) + |p| \leq |y| \quad (\mathrm{I.})$. Setting $y = 0, y = 2p$ respectively yield

$$
\begin{cases} p(x-p) \geq |p| \\ p(x-p) \leq |p| \end{cases} \implies p(x-p) = |p| \qquad (\mathrm{II.})
$$

Therefore $(I.)$ becomes $(\forall y \in \mathbb{R})\ (y-p)(x-p) + p(x-p) \leq |y| \implies (\forall y \in \mathbb{R}^m)\ y(x-p) \leq |y|$. Take $y > 0$ and $y < 0$ yields respectively,

$$
\begin{cases} p \geq x - 1 \\ p \leq x + 1 \end{cases} \qquad (\mathrm{III.})
$$

- If $x > 1$ then $(\mathrm{III.})$ and $(\mathrm{II.})$ combined implies that $p = x - 1$
- If $x < -1$ then $p = x + 1$.
- If $-1 \leq x \leq 1$, left as exercise. □

Proposition L10-h: Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex, l.s.c. and proper, then x minimizes f over $\mathbb{R}^m$ if and only if

$$
x = \mathrm{prox}_f(x)
$$

Proof: let $x \in \mathbb{R}^m$, by proposition L10-g we have $x = \mathrm{prox}_f(x) \iff (\forall y \in \mathbb{R}^m)\langle y-x, x-x\rangle + f(x) \leq f(y) \iff (\forall y \in \mathbb{R}^m)\ f(x) \leq f(y)$. □

More on Proximal Operator

Why prox operators of convex functions are really "nice"? Consider the function f, g, h defined on the real line: given $\lambda > 0$, $f(x) = 0,\ g(x) = \begin{cases} 0, & x \neq 0 \\ -\lambda, & x = 0 \end{cases},\ h(x) = \begin{cases} 0, & x \neq 0 \\ \lambda, & x = 0 \end{cases}$. Clearly, f is convex, but g, h are not convex.

- $\mathrm{prox}_f(x) = x$: let $x \in \mathbb{R}$, $\mathrm{prox}_f(x)$ is the "unique" minimizer of the function $\frac{1}{2}(y-x)^2 \geq 0$
- Finding prox_g: let $x \in \mathbb{R}$. $\mathrm{prox}_g(x)$ is the minimizer of the function $K(y) = g(y) + \frac{1}{2}(y-x)^2 = \begin{cases} \frac{1}{2}(y-x)^2, & y \neq 0 \\ \frac{1}{2}x^2 - \lambda, & y = 0 \end{cases}$. Let K_{opt} be the minimum value of $K(y)$. Observe that if $x^2 \geq 2\lambda$ then $K_{\mathrm{opt}} \geq 0$. If $x^2 > 2\lambda$, then $K_{\mathrm{opt}} = 0$ and is attained if and only if $y = x$. Else if $x^2 = 2\lambda$ then $K_{\mathrm{opt}} = 0$ and is attained if and only if $y \in \{0, x\}$. Lastly, if $x^2 < 2\lambda$ then $K_{\mathrm{opt}} = \frac{1}{2}x^2 - 2\lambda$ and is attained if and only if $y = 0$. Therefore

$$
\mathrm{prox}_g(x) = \begin{cases} \{x\}, & |x| > \sqrt{2\lambda} \\ \{0, x\}, & |x| = \sqrt{2\lambda} \\ \{0\}, & |x| < \sqrt{2\lambda} \end{cases}
$$

- $\mathrm{prox}_h(x) = \begin{cases} \{x\}, & x \neq 0 \\ \emptyset, & x = 0 \end{cases}$, i.e. $\mathrm{prox}_h(x)$ is not defined at $x = 0$. So convexity is critical for the proximal operator to be well defined.

Example L11-1: let $f : \mathbb{R} \to \mathbb{R} : x \mapsto \lambda|x|$ where $\lambda \geq 0$. Then f is convex. We claim that $(\forall x \in \mathbb{R})$,

$$\operatorname{prox}_f(x) = \begin{cases} x - \lambda, & x > \lambda \\ 0, & -\lambda \leq x \leq \lambda \\ x + \lambda, & x < -\lambda \end{cases}$$

This is known as the **soft threshold**.

The above formula is often written as

$$\operatorname{prox}_f(x) = \operatorname{sgn}(x) \max\{|x| - \lambda, 0\}$$

Theorem L11-a: Let $f : \mathbb{R}^m \longrightarrow (-\infty, +\infty]$ be ginve by $(\forall x = (x_1, \ldots, x_m) \in \mathbb{R}^m)\ f(x_1, x_2, \ldots, x_m) = \sum_{i=1}^{m} f_i(x_i)$ where $(\forall i \in \{1, \ldots, m\})\ f_i : \mathbb{R} \to (-\infty, +\infty]$ is convex, l.s.c. and proper. Then $(\forall x \in \mathbb{R}^m)$, we have

$$\begin{aligned} \operatorname{prox}_f(x) &= (\operatorname{prox}_{f_i}(x_i))_{i=1}^m \\ &= (\operatorname{prox}_{f_1}(x_1), \ldots, \operatorname{prox}_{f_m}(x_m)) \end{aligned}$$

Proof: it follows from assignment 2 that f is convex, l.s.c. and proper (f is direct sum). Let $p = (p_1, p_2, \ldots, p_m) \in \mathbb{R}^m$. Then by proposition L10-g we have $p = \operatorname{prox}_f(x)$ if and only if $(\forall y \in \mathbb{R}^m)\ f(y) \geq f(p) + \langle y - p, x - p \rangle$ if and only if $(\forall \{y_1, \ldots y_m\} \subseteq \mathbb{R})\ \sum f_i(y_i) \geq \sum f_i(p_i) + \sum (y_i - p_i)(x_i - p_i)$.

Setting $(\forall i \in \{2, \ldots, m\})\ y_i = p_i$ yields that $(\forall y_1 \in \mathbb{R})\ f_1(y_1) \geq f_1(p_1) + (y_1 - p_1)(x_1 - p_1)$ if and only if $p_1 = \operatorname{prox}_{f_1}(x_1)$. Similar arguments yield that $(\forall i \in \{1, \ldots, m\})\ p_i = \operatorname{prox}_{f_i}(x_i)$, and the proof is complete. □

Example L11-3: Let $g : \mathbb{R}^m \to (-\infty, +\infty]$ be given by $g(x) = \begin{cases} -\alpha \sum_{i=1}^{m} \log x_i, & x > 0 \\ +\infty, & \text{otherwise} \end{cases}$ with $\alpha > 0$. Then

$$\operatorname{prox}_g(x) = \left(\frac{x_i + \sqrt{x_i^2 + 4\alpha}}{2} \right)_{i=1}^m$$

Proof: consider the function $f : \mathbb{R} \to (-\infty, +\infty]$ where $(\forall x \in \mathbb{R})\ f(x) = \begin{cases} -\alpha \log x, & x > 0 \\ +\infty, & \text{otherwise} \end{cases}$. Then f is convex, l.s.c. and proper. Indeed, f is differentiable and thus l.s.c., f is also proper since $f > -\infty, \operatorname{dom}(f) \neq \emptyset$. Moreover, $(\forall x > 0)\ f''(x) = \frac{\alpha}{x^2} > 0$ which implies the function is convex (from Lemma L8-a).

We claim that $(\forall x \in \mathbb{R}), \operatorname{prox}_f(x) = \frac{x + \sqrt{x^2 + 4\alpha}}{2}$. Indeed, recall that $p = \operatorname{prox}_f(x)$ is the unique minimizer of the function $h(y) = f(y) + \frac{1}{2}(y - x)^2 = \begin{cases} -\alpha \log y + \frac{1}{2}(y - x)^2, & y > 0 \\ +\infty, & \text{otherwise} \end{cases}$. Clearly, h is differentiable on its domain $(0, +\infty)$. Therefore, $p = \operatorname{prox}_f(x) \iff h'(p) = 0 \iff (-\alpha \log p + \frac{1}{2}(p - x)^2)' = 0 \iff -\frac{\alpha}{p} + p - x = 0$, since p must be in domain $(p > 0)$, we have $p = \frac{x + \sqrt{x^2 + 4\alpha}}{2}$.

The result then follows from Theorem L11-a. □

Theorem L11-b: Let $g : \mathbb{R}^m \to (-\infty, +\infty]$ be proper, let $c > 0$, let $a \in \mathbb{R}^m$, let $\gamma \in \mathbb{R}$, and define function

$$\begin{aligned} f : \mathbb{R}^m &\longrightarrow \mathbb{R} \\ x &\longmapsto g(x) + \frac{c}{2}\|x\|^2 + \langle a, x \rangle + \gamma \end{aligned}$$

Then $(\forall x \in \mathbb{R}^m)$,

$$\operatorname{prox}_f(x) = \operatorname{prox}_{\frac{1}{c+1} g}\left(\frac{x - a}{c + 1} \right)$$

Proof: Indeed, recall that

$$\begin{aligned} \operatorname{prox}_f(x) &= \underset{v \in \mathbb{R}^m}{\arg\min} \left\{ f(v) + \frac{1}{2}\|u - x\|^2 \right\} \quad \text{assume uniqueness} \\ &= \underset{v \in \mathbb{R}^m}{\arg\min} \left\{ g(v) + \frac{c}{2}\|v\|^2 + \langle a, v \rangle + \gamma + \frac{1}{2}\|v - x\|^2 \right\} \end{aligned}$$

Now,

$$
\begin{aligned}
&\frac{c}{2}\|v\|^2 + \langle a, v\rangle + \frac{1}{2}\|v-x\|^2 \\
&= \frac{c}{2}\|v\|^2 + \langle a, v\rangle + \frac{1}{2}\|v\|^2 - \langle v, x\rangle + \frac{1}{2}\|x\|^2 \\
&= \frac{c+1}{2}\|v\|^2 - \langle v, x-a\rangle + \frac{1}{2}\|x\|^2 \\
&= \frac{c+1}{2}\left[\left\|v - \left(\frac{x-a}{c+1}\right)\right\|^2 - \frac{\|x-a\|^2}{(c+1)^2} + \frac{1}{c+1}\|x\|^2\right]
\end{aligned}
$$

Observe that for any function h, and constants $c \in \mathbb{R}, \alpha > 0$, we have $\underset{v\in\mathbb{R}^m}{\arg\min}\{\alpha h(v) + c\} = \underset{v\in\mathbb{R}^m}{\arg\min}\{h(v)\}$, and so

$$
\begin{aligned}
\operatorname{prox}_f(x) &= \underset{v\in\mathbb{R}^m}{\arg\min}\left\{g(v) + \frac{c+1}{2}\left\|v - \left(\frac{x+a}{c+1}\right)\right\|^2 + \gamma - \frac{\|x-a\|^2}{(c+1)^2} + \frac{1}{c+1}\|x\|^2\right\} \\
&= \underset{v\in\mathbb{R}^m}{\arg\min}\left\{g(v) + \frac{c+1}{2}\left\|v - \left(\frac{x+a}{c+1}\right)\right\|^2\right\} \\
&= \underset{v\in\mathbb{R}^m}{\arg\min}\left\{\frac{1}{c+1}g(v) + \frac{1}{2}\left\|v - \left(\frac{x+a}{c+1}\right)\right\|^2\right\} \\
&= \operatorname{prox}_{\frac{1}{c+1}g}\left(\frac{x+a}{c+1}\right)
\end{aligned}
$$

And we're done. □

Example L11-5: let $\alpha \in [0, +\infty)$ and let $c = [0, \alpha]$, set $f = \delta_C$. Then $(\forall x \in \mathbb{R})\ \operatorname{prox}_f(x) = P_C(x) = \begin{cases} 0, & x \le 0 \\ x, & 0 < x < \alpha \\ \alpha, & x \ge \alpha \end{cases} = \min\{\max\{x, 0\}, \alpha\}$. This follows from Example L10-13.

Example L11-6: Let $f : \mathbb{R} \to (-\infty, +\infty]$ be given by $(\forall x \in \mathbb{R})\ f(x) = \begin{cases} \mu x, & 0 \le x \le \alpha \\ +\infty, & \text{otherwise} \end{cases}$ where $\mu \in \mathbb{R}, \alpha \ge 0$. Then $(\forall x \in \mathbb{R})\ f(x) = \mu x + \delta_{[0,\alpha]}(x)$. Moreover, $\operatorname{prox}_f(x) = \min\{\max\{x - \mu, 0\}, \alpha\}$.

Proof: observe that f is proper, convex, l.s.c. Apply Theorem L11-b with $c = \gamma = 0, g = \delta_{[0,\alpha]}, a = \mu$ and in view of Example L11-5 with $C = [0, \alpha]$ yield:

$$\operatorname{prox}_f(x) = \operatorname{prox}_g(x - \mu) = P_C(x - \mu)$$

and the result follows. □

Theorem L12-a: Let $g : \mathbb{R} \to (-\infty, +\infty]$ be convex, l.s.c. and proper such that $\operatorname{dom}(g) \subseteq [0, +\infty)$ and let $f : \mathbb{R}^m \to \mathbb{R}$ be given by $f(x) = g(\|x\|)$, then

$$\operatorname{prox}_f(x) = \begin{cases} \operatorname{prox}_g(\|x\|)\frac{x}{\|x\|}, & x \ne 0 \\ \{u \in \mathbb{R}^m : \|u\| = \operatorname{prox}_g(0)\}, & x = 0 \end{cases}$$

Proof: In the case $x = 0$, by definition we have $\operatorname{prox}_f(0)$ is the set $\underset{u\in\mathbb{R}^m}{\arg\min}\{f(u) + \frac{1}{2}\|u\|\}$. Using hte change of variable $w = \|u\|$ the above minimizer is the same as $\underset{w\in[0,+\infty)}{\arg\min}\{g(w) + \frac{1}{2}w^2\} = \operatorname{prox}_g(0)$. Hence $\operatorname{prox}_f(0) = \{u \in \mathbb{R}^m : \|u\| = \operatorname{prox}_g(0)\}$.

In the case $x \ne 0$: $\operatorname{prox}_f(x)$ is the set of solutions of the problem

$$
\begin{aligned}
\min_{u\in\mathbb{R}^m}\left\{g(\|u\|) + \frac{1}{2}\|u - x\|\right\} &= \min_{u\in\mathbb{R}^m}\left\{g(\|u\|) + \frac{1}{2}\|u\|^2 - \langle u, x\rangle + \frac{1}{2}\|x\|^2\right\} \\
&= \min_{\alpha\ge 0}\min_{u\in\mathbb{R}^m, \|u\|=\alpha}\left\{g(\alpha) + \frac{1}{2}\alpha^2 - \alpha\|x\|\cos\theta_{u,x} + \frac{1}{2}\|x\|^2\right\} \\
&= \min_{\alpha\ge 0}\left\{g(\alpha) + \frac{1}{2}(\alpha - \|x\|)^2\right\} = \operatorname{prox}_g(x)
\end{aligned}
$$

Observe that $\min_{u\in\mathbb{R}^m, \|u\|=\alpha}(-\langle u, x\rangle) = -\alpha\|x\|$ and is attained at $u = \alpha\frac{x}{\|x\|}$. Hence, $\operatorname{prox}_f(x) = \operatorname{prox}_g(x)\frac{x}{\|x\|}$. □

Example L12-2: follow-up of Example L11-6, let $\alpha > 0$, let $f : \mathbb{R} \to (-\infty, +\infty]$ be given by $(\forall x \in \mathbb{R})\ f(x) = \begin{cases} \lambda|x|, & |x| \leq \alpha \\ +\infty, & \text{otherwise} \end{cases}$ where $\lambda \geq 0$. Then f is convex, l.s.c. and proper. Moreover $(\forall x \in \mathbb{R})\ \text{prox}_f(x) = \min\{\max\{|x| - \lambda, 0\}, \alpha\} \cdot \text{sgn}(x)$.

Proof: define $(\forall x \in \mathbb{R})\ g(x) = \begin{cases} \lambda x, & 0 \leq x \leq \alpha, \\ +\infty, & \text{otherwise} \end{cases}$, moreover, $\text{dom}(g) = [0, \alpha] \subseteq [0, +\infty)$. Moreover, $(\forall x \in \mathbb{R})\ f(x) = g(|x|)$. Using Example L11-6, we have $\text{prox}_g(0) = \min\{\max\{-\lambda, 0\}, \alpha\} = 0$. Hence by Theorem 12-a, $\text{prox}_f(x) = \begin{cases} \text{prox}_g(|x|)\text{sgn}(x), & x \neq 0 \\ 0 & x = 0 \end{cases}$ which is equivalent to the result claimed. □

Example L12-3: let $w = (w_1, \ldots, w_m) \in \mathbb{R}^m_+$, let $\alpha = (\alpha_1, \ldots, \alpha_m) \in \mathbb{R}^m_+$, let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be given by

$$f(x) = \begin{cases} \sum_{i=1}^m w_i |x_i|, & -\alpha \leq x \leq \alpha \\ +\infty, & \text{otherwise} \end{cases}$$

Then,

1. $\text{prox}(x) = (\min\{\max\{|x_i| - w_i, 0\}, \alpha_i) \cdot \text{sgn}(x_i))_{i=1}^m$
2. Let $x_0 \in \mathbb{R}^m$. $(\forall n \in \mathbb{N})$ update via $x_{n+1} = \text{prox}_f(x_n)$. Then $x_n \to \bar{x}$ where $\bar{x}$ solves the problem,

$$\min \left(\sum_{i=1}^m w_i |x_i| : |x_i| \leq \alpha_i \forall i \in \{1, \ldots, m\} \right)$$

Nonexpansive, Firmly Nonexpansive, and Averaged Operators

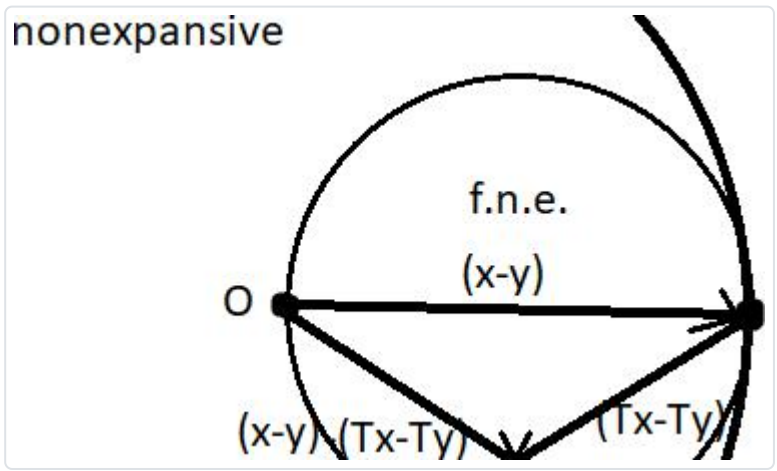
From now on, we shall use Id to denote the $m \times m$ identity matrix on $\mathbb{R}^m$, i.e. $\mathrm{Id} : \mathbb{R}^m \to \mathbb{R}^m : x \mapsto x$.

Nonexpansiveness: Let $T : \mathbb{R}^m \to \mathbb{R}^m$. Then T is nonexpansive if $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$, $\|Tx - Ty\| \leq \|x - y\|$.

> Nonexpansive operators are a special case of Lipschitz continuous with constant at most 1.

T is firmly nonexpansive if $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$, $\|Tx - Ty\|^2 + \|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 \leq \|x - y\|^2$.

α-**Average**: Let $\alpha \in (0, 1)$. Then T is α-averaged if $(\exists N : \mathbb{R}^m \to \mathbb{R}^m)$ N is nonexppansive and $T = (1 - \alpha)\mathrm{Id} + \alpha N$.

> Firm Nonexpansive (f.n.e.) $\implies$ Averaged $\implies$ nonexpansiveness, see next Remark L12-7.
>
> For visualization
>
> 
>

Proposition L12-b: let $T : \mathbb{R}^m \to \mathbb{R}^m$. Then the following are equivalent:

1. T is f.n.e.
2. $\mathrm{Id} - T$ is f.n.e.
3. $2T - \mathrm{Id}$ is nonexpansive
4. $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$ $\|Tx - Ty\|^2 \leq \langle x - y, Tx - Ty \rangle$
5. $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$ $\langle Tx - Ty, (\mathrm{Id} - T)x - (\mathrm{Id} - T)y \rangle \geq 0$

Proof: $(1. \iff 2.)$ clearly follows from the definition. $(1. \iff 3. \iff 4. \iff 5.)$ are in assignment 3.

Proposition L12-c: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be linear. Then the following are equivalent:

1. T is f.n.e.
2. $\|2T - \mathrm{Id}\| \leq 1$
3. $(\forall x \in \mathbb{R}^m)$ $\|Tx\|^2 \leq \langle x, Tx \rangle$
4. $(\forall x \in \mathbb{R}^m)$ $\langle Tx, x - Tx \rangle \geq 0$

Proof: $(1. \iff 2.)$ using Proposition L12-b we have T is f.n.e. if and only if $2T - \mathrm{Id}$ is nonexpansive. Hence $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$ $\|(2T - \mathrm{Id})x - (2T - \mathrm{Id})y\| \leq \|x - y\|$ and by linearity this is equivalent to $(\forall z \in \mathbb{R}^m)$ $\|(2T - \mathrm{Id})z\| \leq \|z\|$, this implies that $(\forall z \in \mathbb{R}^m \backslash \{0\}) \frac{\|(2T - \mathrm{Id})z\|}{\|z\|} \leq 1 \implies \sup \frac{\|(2T - \mathrm{Id})z\|}{\|z\|} \leq 1 \implies \|2T - \mathrm{Id}\| \leq 1$.

Conversely, suppose that $\|2T - \mathrm{Id}\| \leq 1$. Then $(\forall z \in \mathbb{R}^m \backslash \{0\}) \frac{\|(2T - \mathrm{Id})z\|}{\|z\|} \leq 1 \implies (\forall z \in \mathbb{R}^m) \|(2T - \mathrm{Id})z\| \leq \|z\|$. Let $x \in \mathbb{R}^m, y \in \mathbb{R}^m$, setting $z = x - y$ yields the desired result.

Part 3 and 4 are direct consequences of linearity from proposition L12-b. □

Remark L12-7: it follows from the equivalence T is f.n.e. if and only if T is $\frac{1}{2}$-averaged.

Proof: indeed, $2T - \mathrm{Id} = N$ is nonexpansive and $T = \frac{1}{2}\mathrm{Id} + \frac{1}{2}N$.

Example L12-8: Let C be convex, closed and nonempty subset of $\mathbb{R}^m$. Then P_C is f.n.e.

Proof: from Example L10-1 we know that $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)$ $\|P_C(x) - P_C(y)\|^2 \leq \langle P_C(x) - P_C(y), x - y \rangle$. The result follows from point 4 of proposition L12-b. □

Example L12-9: the converse of Remark L12-7 is NOT true. Suppose that $T = -\mathrm{Id}$. Then T in averaged but not f.n.e. Indeed, $T = \frac{1}{4}\mathrm{Id} + \frac{3}{4}(-\mathrm{Id}) \implies T$ is $\frac{3}{4}$-averaged. T is not f.n.e. because $(\forall x \in \mathbb{R}^m)$, $\|Tx\|^2 + \|x - Tx\|^2 = \frac{5}{2}\|x\|^2 > \|x\|^2$ whenever $x \neq 0$.

Example L12-10: suppose that $T = -\mathrm{Id}$. Then T is nonexpansive, but T is NOT averaged. Indeed, suppose for a contradiction $(\exists \alpha \in (0,1))$ and exists nonexpansive $N : \mathbb{R}^m \to \mathbb{R}^m$ such that $T = (1-\alpha)\mathrm{Id} + \alpha N$. Then $(-2+\alpha)\mathrm{Id} = \alpha N$ and $N = \frac{\alpha-2}{\alpha}\mathrm{Id}$. Observe that N is non expansive if and only if $|\frac{\alpha-2}{\alpha}| \leq 1$, which leads to $\alpha > 1$. This contradicts the assumption.

Proposition L12-d: let $T : \mathbb{R}^m \to \mathbb{R}^m$ be nonexpansive, then T is continuous.

Proof: let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$ such that $x_n \to \bar{x}$. We want to show that $T(x_n) \to T(\bar{x})$. Indeed, $(\forall n \in \mathbb{N})\ 0 \leq \|T(x_n) - T(\bar{x})\| \leq \|x_n - \bar{x}\|$. Letting $n \to \infty$, $0 \leq \| \lim_{n\to\infty} T(x_n) - T(\bar{x})\| \leq 0$. Therefore $T(x_n) \to T(\bar{x})$ as claimed. □

Fixed Points: Let $T : \mathbb{R}^m \to \mathbb{R}^m$. Then

$$\mathrm{Fix}(T) = \{x \in \mathbb{R}^m : x = Tx\}$$

> Tip: recall from Proposition L10-h that for convex l.s.c. proper functions $\mathrm{prox}_f(x) = x$ if and only if x minimizes f. In this case x is the unique $\mathrm{Fix}(\mathrm{prox}_f)$.

Fejér Monotonicity: Let C be a nonempty subset of $\mathbb{R}^m$ and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$. Then $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C if $(\forall c \in C)(\forall n \in \mathbb{N})\ \|x_{n+1} - c\| \leq \|x_n - c\|$.

Example L13-2: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be nonexpansive, with $\mathrm{Fix}(T) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m$, $(\forall n \in \mathbb{N})$ update via $x_{n+1} = T(x_n)$. Then $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$.

Indeed, $(\forall f \in \mathrm{Fix}(T))$

$$\begin{aligned}\|x_{n+1} - f\| &= \|T^n(x_0) - T^n(f)\| \\ &= \|T(T^{n-1}(x_0)) - T(T^{n-1}(f))\| \\ &\leq \|T^{n-1}(x_0) - T^{n-1}(f)\| \\ &= \|x_n - f\|\end{aligned}$$

And the result follows with a quick induction.

Proposition L13-a: Let $\phi \neq C \subseteq \mathbb{R}^m$, let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$. Suppose $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C. Then the following hold:

1. $(x_n)_{n\in\mathbb{N}}$ is bounded.
2. $(\forall c \in C)$ the sequence $(\|x_n - c\|)_{n\in\mathbb{N}}$ converges.
3. $(d_C(x_n))_{n\in\mathbb{N}}$ is decreasing and converges.

Proof: For part 1, let $c \in C$. By the triangle inequality $(\forall n \in \mathbb{N})$, we have

$$\begin{aligned}\|x_n\| &\leq \|c\| + \|x_n - c\| \\ &\leq \|c\| + \|x_{n-1} - c\| \\ &\vdots \\ &\leq \|c\| + \|x_0 - c\|\end{aligned}$$

Hence, $(x_n)_{n\in\mathbb{N}}$ is bounded as claimed.

For part 2, observe that $(\forall n \in \mathbb{N})(\forall c \in C)\ 0 \leq \|x_{n+1} - c\| \leq \|x_n - c\|$. From real analysis, a nonincreasing sequence of real numbers bounded below implies that the sequence converges. Part 3 uses a similar approach. □

Lemma L13-b: Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$ and let $c \neq \emptyset$ subset of $\mathbb{R}^m$. Suppose that for every $c \in C$, $(\|x_n - c\|)_{n\in\mathbb{N}}$ converges and that every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in C. Then $(x_n)_{n\in\mathbb{N}}$ converges to a point in C.

Proof: observe that $(x_n)_{n\in\mathbb{N}}$ is bounded, since $\|x_n\| \leq \underbrace{\|x_n - c\|}_{\text{convergent}} + \underbrace{\|c\|}_{\text{const.}}$. Let x, y be two cluster points of $(x_n)_{n\in\mathbb{N}}$. That is $x_{k_n} \to x, y_{l_n} \to y$. By assumption $x \in C, y \in C$.

Observe that $\underbrace{\|x_n - y\|^2}_{\text{converge}} - \underbrace{\|x_n - x\|^2}_{\text{converge}} + \|x\|^2 - \|y\|^2 = 2\langle x_n, x - y\rangle$ which converges, say to l. Taking hte limit along x_{k_n} and x_{l_n} respectively yield $\langle x, x-y\rangle = \langle y, x-y\rangle = l$, which implies $\|x-y\|^2 = \langle x, x-y\rangle - \langle y, x-y\rangle = 0 \implies x = y$. □

Theorem L13-c: Let $\emptyset \neq C \subseteq \mathbb{R}^m$ and let (x_n) be a sequence in $\mathbb{R}^m$. Suppose that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C, and that every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in C. Then $(x_n)_{n\in\mathbb{N}}$ converges to a point in C.

Proof: follows from Proposition L13-a point 2 and Lemma L13-b. □

Theorem L13-d: Let $\alpha \in (0,1)$ and let $T : \mathbb{R}^m \to \mathbb{R}^m$ be α-averaged, such that $\mathrm{Fix}(T) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m$. Update via $(\forall n \in \mathbb{N})\ x_{n+1} = T(x_n)$. Then the following hold:

1. $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$.
2. $T(x_n) - x_n \longrightarrow 0$.
3. $(x_n)_{n\in\mathbb{N}}$ converges to a point in $\mathrm{Fix}(T)$

> Useful Identity for proof: let $x \in \mathbb{R}^m$, let $y \in \mathbb{R}^m$ and let $\alpha \in \mathbb{R}$. One could directly verify that $\|\alpha x + (1-\alpha)y\|^2 + \alpha(1-\alpha)\|x-y\|^2 = \alpha\|x\|^2 + (1-\alpha)\|y\|^2$.

Proof: For part 1, T is averaged implies that T is nonexpansive, and the result follows from Example L13-2.

For part 2. By assumption, $\exists N : \mathbb{R}^m \to \mathbb{R}^m$ such that N is nonexpansive and $T = (1-\alpha)\mathrm{Id} + \alpha N$. Hence $(\forall n \in \mathbb{N})\ x_{n+1} = T(x_n) = (1-\alpha)x_n + \alpha N(x_n)$. Now let $f \in \mathrm{Fix}(T)$.

$$
\begin{aligned}
&\|x_{n+1} - f\|^2 \\
&= \|(1-\alpha)x_n + \alpha N(x_n) - f\|^2 \\
&= (1-\alpha)\|x_n - f\|^2 + \alpha\|N(x_n) - N(f)\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \\
&\leq (1-\alpha)\|x_n - f\|^2 + \alpha\|x_n - f\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \\
&= \|x_n - f\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2
\end{aligned}
$$

Telescoping yields that $\sum_{n=0}^{\infty} \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \leq \|x_0 - f\|^2 - 0 < +\infty$. That is, $\|N(x_n) - x_n\| \to 0$. Recall that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$. Observe also that $\mathrm{Fix}(T) = \mathrm{Fix}(N)$. Altogether, we learn that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(N)$.

For part 3. Let $\bar{x}$ be a cluster point of $(x_n)_{n\in\mathbb{N}}$ say $x_{k_n} \to \bar{x}$. Observe that N is nonexpansive $\implies$ N is continuous. Now recall that $Nx_n - x_n \to 0$. Taking the limit along the subsequence x_{k_n} we learn that $N\bar{x} - \bar{x} = 0$. That is, every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in $\mathrm{Fix}(N) = \mathrm{Fix}(T)$. Now combine with Theorem L13-c, and we're done. □

Corollary L14-a: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be f.n.e. and suppose that $\mathrm{Fix}(T) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m,\ (\forall n \in \mathbb{N})$ update via $x_{n+1} = T(x_n)$. Then $\exists \bar{x} \in \mathrm{Fix}(T)$ such that $x_n \to \bar{x}$.

Proof: f.n.e. $\implies$ α-averaged. The result follows from Theorem L13-d part 3. □

Proposition L14-b: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Then prox_f is f.n.e.

Proof: Let $x, y \in \mathbb{R}^m$. Set $p = \mathrm{prox}_f(x), q = \mathrm{prox}_f(y)$. Using Proposition L10-g, (generalized characterization of projection), we have $(\forall z \in \mathbb{R}^m)$

$$
\begin{cases} \langle z-p, x-p\rangle + f(p) \leq f(z) \\ \langle z-q, y-q\rangle + f(q) \leq f(z) \end{cases} \implies \begin{cases} \langle q-p, x-p\rangle + f(p) \leq f(q), & z = q \\ \langle p-q, y-q\rangle + f(q) \leq f(p), & z = p \end{cases}
$$

Adding the two inequalities yields $\langle q-p, (x-p)-(y-q)\rangle \leq 0$ and equivalently $\langle p-q, (x-p)-(y-q)\rangle \geq 0$. Recall from A3Q3 (i) we know $T : \mathbb{R}^m \to \mathbb{R}^m$ is f.n.e. if and only if $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \langle (\mathrm{Id}-T)(x) - (\mathrm{Id}-T)(y), T(x) - T(y)\rangle \geq 0$. This yields the desired conclusion. □

Corollary L14-c: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex l.s.c. and proper, such that $\arg\min(f) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m$. $(\forall n \in \mathbb{N})$ update via $x_{n+1} = \mathrm{prox}_f(x_n)$. Then $\exists \bar{x} \in \arg\min f$ such that $x_n \to \bar{x}$.

Proof: Observe that by Proposition L10-h, $\arg\min f = \mathrm{Fix}(\mathrm{prox}_f) \neq \emptyset$. Recall that prox_f is f.n.e. by Proposition L14-b, now combine with Corollary L14-a, we arrive at the result. □

Proposition L14-d: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be nonexpansive and let $\alpha \in (0,1)$. Then the following are equivalent:

1. T is α-averaged
2. $(1-\frac{1}{\alpha})\mathrm{Id} + \frac{1}{\alpha}T$ is nonexpansive
3. $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \|T(x) - T(y)\|^2 \leq \|x-y\|^2 - \frac{1-\alpha}{\alpha}\|(\mathrm{Id}-T)(x) - (\mathrm{Id}-T)(y)\|^2$

> Useful Identity for proof: Let $x, y \in \mathbb{R}^m, \alpha \in \mathbb{R}\backslash\{0\}$. Then $\alpha^2(\|x\|^2 - \|(1-\frac{1}{\alpha})x + \frac{1}{\alpha}y\|^2) = \alpha(\|x\|^2 - \frac{1-\alpha}{\alpha}\|x-y\|^2 - \|y\|^2)$.

Proof: $(1. \iff 2.)$ follows from the definition.

$(2. \iff 3.)$ makes use of the above identity, and also follows from the definition. $\square$

Proposition L14-e: Let $\alpha_1, \alpha 2 \in (0,1)$, let $T_i : \mathbb{R}^m \to \mathbb{R}^m$ be α_i-averaged (where i can be 1 or 2). Set $T \coloneqq T_1 \circ T_2, \alpha \coloneqq \frac{\alpha_1+\alpha_2-2\alpha_1\alpha_2}{1-\alpha_1\alpha_2}$. Then T is α-averaged.

Proof Sketch: First show $\alpha \in (0,1)$. Then

$$\begin{aligned}
\|T(x) - T(y)\|^2 &= \|T_1 \circ T_2(x) - T_1 \circ T_2(y)\|^2 \\
&\leq \|T_2(x) - T_2(y)\|^2 - \frac{1-\alpha_1}{\alpha_1}\|(\mathrm{Id} - T_1)(T_2(x)) - (\mathrm{Id} - T_1)(T_2(y))\|^2 \\
&\leq \|x-y\|^2 - \underbrace{\frac{1-\alpha_2}{\alpha_2}\|(\mathrm{Id} - T_2)(x) - (\mathrm{Id} - T_2)(y)\|^2}_{\text{(I.)}} \\
&\qquad - \underbrace{\frac{1-\alpha_1}{\alpha_1}\|(\mathrm{Id} - T_1)(T_2(x)) - (\mathrm{Id} - T_1)(T_2(y))\|^2}_{\text{(II.)}}
\end{aligned}$$

Set $\beta = \frac{1-\alpha_1}{\alpha_1} + \frac{1-\alpha_2}{\alpha_2} > 0$. Show that $\text{(I.)} + \text{(II.)} \geq \frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2}\|(\mathrm{Id} - T)(x) - (\mathrm{Id} - T)(y)\|^2$. Consequently,

$$\|T(x) - T(y)\|^2 \leq \|x-y\|^2 - \frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2}\|(\mathrm{Id} - T)(x) - (\mathrm{Id} - T)(y)\|^2$$

Finally, we verify that $\frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2} = \frac{1-\alpha}{\alpha}$, and the results follow from Proposition L14-d. $\square$

Fejér Monotonicity: Let C be a nonempty subset of $\mathbb{R}^m$ and let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$. Then $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C if $(\forall c \in C)(\forall n \in \mathbb{N})\ \|x_{n+1} - c\| \leq \|x_n - c\|$.

Example L13-2: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be nonexpansive, with $\mathrm{Fix}(T) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m$, $(\forall n \in \mathbb{N})$ update via $x_{n+1} = T(x_n)$. Then $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$.

Indeed, $(\forall f \in \mathrm{Fix}(T))$

$$\begin{aligned}
\|x_{n+1} - f\| &= \|T^n(x_0) - T^n(f)\| \\
&= \|T(T^{n-1}(x_0)) - T(T^{n-1}(f))\| \\
&\leq \|T^{n-1}(x_0) - T^{n-1}(f)\| \\
&= \|x_n - f\|
\end{aligned}$$

And the result follows with a quick induction.

Proposition L13-a: Let $\phi \neq C \subseteq \mathbb{R}^m$, let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$. Suppose $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C. Then the following hold:

1. $(x_n)_{n\in\mathbb{N}}$ is bounded.
2. $(\forall c \in C)$ the sequence $(\|x_n - c\|)_{n\in\mathbb{N}}$ converges.
3. $(d_C(x_n))_{n\in\mathbb{N}}$ is decreasing and converges.

Proof: For part 1, let $c \in C$. By the triangle inequality $(\forall n \in \mathbb{N})$, we have

$$\begin{aligned}
\|x_n\| &\leq \|c\| + \|x_n - c\| \\
&\leq \|c\| + \|x_{n-1} - c\| \\
&\vdots \\
&\leq \|c\| + \|x_0 - c\|
\end{aligned}$$

Hence, $(x_n)_{n\in\mathbb{N}}$ is bounded as claimed.

For part 2, observe that $(\forall n \in \mathbb{N})(\forall c \in C)\ 0 \leq \|x_{n+1} - c\| \leq \|x_n - c\|$. From real analysis, a nonincreasing sequence of real numbers bounded below implies that the sequence converges. Part 3 uses a similar approach. $\square$

Lemma L13-b: Let $(x_n)_{n\in\mathbb{N}}$ be a sequence in $\mathbb{R}^m$ and let $c \neq \emptyset$ subset of $\mathbb{R}^m$. Suppose that for every $c \in C$, $(\|x_n - c\|)_{n\in\mathbb{N}}$ converges and that every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in C. Then $(x_n)_{n\in\mathbb{N}}$ converges to a point in C.

Proof: observe that $(x_n)_{n\in\mathbb{N}}$ is bounded, since $\|x_n\| \leq \underbrace{\|x_n - c\|}_{\text{convergent}} + \underbrace{\|c\|}_{\text{const.}}$. Let x, y be two cluster points of $(x_n)_{n\in\mathbb{N}}$. That is $x_{k_n} \to x, y_{l_n} \to y$. By assumption $x \in C, y \in C$.

Observe that $\underbrace{\|x_n - y\|^2}_{\text{converge}} - \underbrace{\|x_n - x\|^2}_{\text{converge}} + \|x\|^2 - \|y\|^2 = 2\langle x_n, x - y\rangle$ which converges, say to l. Taking hte limit along x_{k_n} and x_{l_n} respectively yield $\langle x, x-y\rangle = \langle y, x-y\rangle = l$, which implies $\|x-y\|^2 = \langle x, x-y\rangle - \langle y, x-y\rangle = 0 \implies x = y$. □

Theorem L13-c: Let $\emptyset \neq C \subseteq \mathbb{R}^m$ and let (x_n) be a sequence in $\mathbb{R}^m$. Suppose that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to C, and that every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in C. Then $(x_n)_{n\in\mathbb{N}}$ converges to a point in C.

Proof: follows from Proposition L13-a point 2 and Lemma L13-b. □

Theorem L13-d: Let $\alpha \in (0,1)$ and let $T : \mathbb{R}^m \to \mathbb{R}^m$ be α-averaged, such that $\mathrm{Fix}(T) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m$. Update via $(\forall n \in \mathbb{N})\ x_{n+1} = T(x_n)$. Then the following hold:

1. $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$.
2. $T(x_n) - x_n \longrightarrow 0$.
3. $(x_n)_{n\in\mathbb{N}}$ converges to a point in $\mathrm{Fix}(T)$

> Useful Identity for proof: let $x \in \mathbb{R}^m$, let $y \in \mathbb{R}^m$ and let $\alpha \in \mathbb{R}$. One could directly verify that $\|\alpha x + (1-\alpha)y\|^2 + \alpha(1-\alpha)\|x-y\|^2 = \alpha\|x\|^2 + (1-\alpha)\|y\|^2$.

Proof: For part 1, T is averaged implies that T is nonexpansive, and the result follows from Example L13-2.

For part 2. By assumption, $\exists N : \mathbb{R}^m \to \mathbb{R}^m$ such that N is nonexpansive and $T = (1-\alpha)\mathrm{Id} + \alpha N$. Hence $(\forall n \in \mathbb{N})\ x_{n+1} = T(x_n) = (1-\alpha)x_n + \alpha N(x_n)$. Now let $f \in \mathrm{Fix}(T)$.

$$
\begin{aligned}
&\|x_{n+1} - f\|^2 \\
&= \|(1-\alpha)x_n + \alpha N(x_n) - f\|^2 \\
&= (1-\alpha)\|x_n - f\|^2 + \alpha\|N(x_n) - N(f)\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \\
&\leq (1-\alpha)\|x_n - f\|^2 + \alpha\|x_n - f\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \\
&= \|x_n - f\|^2 - \alpha(1-\alpha)\|N(x_n) - x_n\|^2
\end{aligned}
$$

Telescoping yields that $\sum_{n=0}^{\infty} \alpha(1-\alpha)\|N(x_n) - x_n\|^2 \leq \|x_0 - f\|^2 - 0 < +\infty$. That is, $\|N(x_n) - x_n\| \to 0$. Recall that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(T)$. Observe also that $\mathrm{Fix}(T) = \mathrm{Fix}(N)$. Altogether, we learn that $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}(N)$.

For part 3. Let $\bar{x}$ be a cluster point of $(x_n)_{n\in\mathbb{N}}$ say $x_{k_n} \to \bar{x}$. Observe that N is nonexpansive $\implies$ N is continuous. Now recall that $Nx_n - x_n \to 0$. Taking the limit along the subsequence x_{k_n} we learn that $N\bar{x} - \bar{x} = 0$. That is, every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in $\mathrm{Fix}(N) = \mathrm{Fix}(T)$. Now combine with Theorem L13-c, and we're done. □

Corollary L14-a: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be f.n.e. and suppose that $\mathrm{Fix}(T) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m,\ (\forall n \in \mathbb{N})$ update via $x_{n+1} = T(x_n)$. Then $\exists \bar{x} \in \mathrm{Fix}(T)$ such that $x_n \to \bar{x}$.

Proof: f.n.e. $\implies$ α-averaged. The result follows from Theorem L13-d part 3. □

Proposition L14-b: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Then prox_f is f.n.e.

Proof: Let $x, y \in \mathbb{R}^m$. Set $p = \mathrm{prox}_f(x), q = \mathrm{prox}_f(y)$. Using Proposition L10-g, (generalized characterization of projection), we have $(\forall z \in \mathbb{R}^m)$

$$
\begin{cases} \langle z - p, x - p\rangle + f(p) \leq f(z) \\ \langle z - q, y - q\rangle + f(q) \leq f(z) \end{cases} \implies \begin{cases} \langle q - p, x - p\rangle + f(p) \leq f(q), & z = q \\ \langle p - q, y - q\rangle + f(q) \leq f(p), & z = p \end{cases}
$$

Adding the two inequalities yields $\langle q - p, (x-p) - (y-q)\rangle \leq 0$ and equivalently $\langle p - q, (x-p) - (y-q)\rangle \geq 0$. Recall from A3Q3 (i) we know $T : \mathbb{R}^m \to \mathbb{R}^m$ is f.n.e. if and only if $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \langle (\mathrm{Id} - T)(x) - (\mathrm{Id} - T)(y), T(x) - T(y)\rangle \geq 0$. This yields the desired conclusion. □

Corollary L14-c: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex l.s.c. and proper, such that $\arg\min(f) \neq \emptyset$. Let $x_0 \in \mathbb{R}^m$. $(\forall n \in \mathbb{N})$ update via $x_{n+1} = \mathrm{prox}_f(x_n)$. Then $\exists \bar{x} \in \arg\min f$ such that $x_n \to \bar{x}$.

Proof: Observe that by Proposition L10-h, $\arg\min f = \mathrm{Fix}(\mathrm{prox}_f) \neq \emptyset$. Recall that prox_f is f.n.e. by Proposition L14-b, now combine with Corollary L14-a, we arrive at the result. □

Proposition L14-d: Let $T : \mathbb{R}^m \to \mathbb{R}^m$ be nonexpansive and let $\alpha \in (0,1)$. Then the following are equivalent:

1. T is α-averaged

2. $(1-\frac{1}{\alpha})\mathrm{Id}+\frac{1}{\alpha}T$ is nonexpansive
3. $(\forall x\in\mathbb{R}^m)(\forall y\in\mathbb{R}^m)\ \|T(x)-T(y)\|^2\leq\|x-y\|^2-\frac{1-\alpha}{\alpha}\|(\mathrm{Id}-T)(x)-(\mathrm{Id}-T)(y)\|^2$

> Useful Identity for proof: Let $x,y\in\mathbb{R}^m,\alpha\in\mathbb{R}\backslash\{0\}$. Then $\alpha^2(\|x\|^2-\left\|(1-\frac{1}{\alpha})x+\frac{1}{\alpha}y\right\|^2)=\alpha(\|x\|^2-\frac{1-\alpha}{\alpha}\|x-y\|^2-\|y\|^2)$.

Proof: $(1.\iff 2.)$ follows from the definition.

$(2.\iff 3.)$ makes use of the above identity, and also follows from the definition. □

Proposition L14-e: Let $\alpha_1,\alpha 2\in(0,1)$, let $T_i:\mathbb{R}^m\to\mathbb{R}^m$ be α_i-averaged (where i can be 1 or 2). Set $T:=T_1\circ T_2,\alpha:=\frac{\alpha_1+\alpha_2-2\alpha_1\alpha_2}{1-\alpha_1\alpha_2}$. Then T is α-averaged.

Proof Sketch: First show $\alpha\in(0,1)$. Then

$$\begin{aligned}\|T(x)-T(y)\|^2&=\|T_1\circ T_2(x)-T_1\circ T_2(y)\|^2\\&\leq\|T_2(x)-T_2(y)\|^2-\frac{1-\alpha_1}{\alpha_1}\|(\mathrm{Id}-T_1)(T_2(x))-(\mathrm{Id}-T_1)(T_2(y))\|^2\\&\leq\|x-y\|^2-\underbrace{\frac{1-\alpha_2}{\alpha_2}\|(\mathrm{Id}-T_2)(x)-(\mathrm{Id}-T_2)(y)\|^2}_{\text{(I.)}}\\&\quad\underbrace{-\frac{1-\alpha_1}{\alpha_1}\|(\mathrm{Id}-T_1)(T_2(x))-(\mathrm{Id}-T_1)(T_2(y))\|^2}_{\text{(II.)}}\end{aligned}$$

Set $\beta=\frac{1-\alpha_1}{\alpha_1}+\frac{1-\alpha_2}{\alpha_2}>0$. Show that $(\mathrm{I.})+(\mathrm{II.})\geq\frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2}\|(\mathrm{Id}-T)(x)-(\mathrm{Id}-T)(y)\|^2$. Consequently,

$$\|T(x)-T(y)\|^2\leq\|x-y\|^2-\frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2}\|(\mathrm{Id}-T)(x)-(\mathrm{Id}-T)(y)\|^2$$

Finally, we verify that $\frac{(1-\alpha_1)(1-\alpha_2)}{\beta\alpha_1\alpha_2}=\frac{1-\alpha}{\alpha}$, and the results follow from Proposition L14-d. □

Constrained Convex Optimization

We now consider the problem

$$(\mathrm{P}) \quad \min(f(x) : x \in C)$$

where $f : \mathbb{R}^m \to (-\infty, +\infty]$ convex, l.s.c. and proper; $C \neq \emptyset$ is convex and closed.

Recall Fact L7-e, let $\bar{x} \in \mathbb{R}^m$. Then $\bar{x}$ solves (P) if and only if $(\partial f(\bar{x})) \cap (-N_C(\bar{x})) \neq \emptyset$.

We shall now see some weaker results, in the absence of convexity.

Theorem L15-a: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be proper; Let $g : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. $\mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f))$. Consider the problem:

$$\min_{x \in \mathbb{R}^m} f(x) + g(x)$$

1. If $x^* \in \mathrm{dom}(g)$ is a local optimal of (P) and f is differentiable at x^*, then

$$-\nabla f(x^*) \in \partial g(x^*)$$

2. Suppose that f is convex. If f is differentiable at $x^* \in \mathrm{dom}(g)$ then x^* is a global minimizer of (P) if and only if

$$-\nabla f(x^*) \in \partial g(x^*)$$

Proof: For Part 1, let $y \in \mathrm{dom}(g)$. Since g is convex, we know that $\mathrm{dom}(g)$ is convex. Hence for any $\lambda \in (0,1)$, $x_\lambda \coloneqq x^* + \lambda(y - x^*) = (1-\lambda)x^* + \lambda y \in \mathrm{dom}(g)$. Therefore, for sufficiently small λ, $f(x_\lambda) + g(x_\lambda) \geq f(x^*) + g(x^*)$. By convexity of g we learn that $f((1-\lambda)x^* + \lambda y) + (1-\lambda)g(x^*) + \lambda g(y) \geq f(x^*) + g(x^*)$. Rearranging yields $\lambda g(x^*) - \lambda g(y) \leq f((1-\lambda)x^* + \lambda y) - f(x^*)$, equivalently, $g(x^*) - g(y) \leq \frac{f((1-\lambda)x^* + \lambda y) - f(x^*)}{\lambda}$. Taking the limit as $\lambda \to 0^+$, we obtain $g(y) \geq g(x^*) + \langle -\nabla f(x^*), y - x^* \rangle \implies -\nabla f(x^*) \in \partial g(x^*)$, by definition of subdifferential.

For part 2, suppose that f is convex. Observe that part 1 prove $(\Leftarrow)$. Now suppose that $-\nabla f(x^*) \in \partial g(x^*)$. On the one hand, for any $y \in \mathrm{dom}(g), g(y) \geq g(x^*) + \langle -\nabla f(x^*), y - x \rangle \quad (\mathrm{I.})$. On the other hand, since f is convex, differentiable at x^*, then for any $y \in \mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f)), f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle \quad (\mathrm{II.})$, from Proposition L8-b part 2. Adding $(\mathrm{I.})$ and $(\mathrm{II.})$ yields for any $y \in \mathrm{dom}(g), f(y) + g(y) \geq f(x^*) + g(x^*)$. That is, x^* is optimal solution of (P). $\square$

The KKT Conditions

KKT stands for **Korush-Kuhn-Tucker**.

In the following section we assume $f, g_1, \ldots, g_n$ are functions from $\mathbb{R}^m \to \mathbb{R}$ (full domain!). $I = \{1, \ldots, n\}$. Consider the problem

$$(\mathrm{P}) \quad \min(f(x) : (\forall i \in I)\ g_i(x) \leq 0)$$

We assume that (P) has at least one sollution and that $\mu \coloneqq \min(f(x) : (\forall i \in I)\ g_i(x) \leq 0) \in \mathbb{R}$ be the optimal value. Define

$$F(x) \coloneqq \max\{\underbrace{f(x) - \mu}_{=: g_0(x)}, g_1(x), \ldots, g_n(x)\}$$

Lemma 15-b: we have $(\forall x \in \mathbb{R}^m)\ F(x) \geq 0$. Moreover, solutions of $(\mathrm{P}) = \text{minimizer of } F = \{x : F(x) = 0\}$.

Fact L15-c, max rule for subdifferential calculus: Let $g_1, \ldots, g_n : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Define $g(x) = \max\{g_1(x), \ldots, g_n(x)\}$; $A(x) = \{i \in \{1, \ldots, n\} : g_i(x) = g(x)\}$. Let $x \in \bigcap_{i=1}^{n} \mathrm{int}(\mathrm{dom}(g_i))$. Then

$$\partial g(x) = \mathrm{conv}\left(\bigcup_{i \in A(x)} \partial g_i(x) \right)$$

Theorem L15-d, Fritz-John necessary optimality conditions: Suppose that $f, g_1, \ldots, g_n$ are convex and x^* solves (P). Then $\exists \alpha_0 \geq 0, \ldots, \alpha_n \geq 0$, not all 0, for which

$$0 \in \alpha_0\ \partial f(x^*) + \sum_{i \in I} \alpha_i\ \partial g_i(x^*)$$

And $(\forall i \in I = \{1, \ldots, n\})\ \alpha_i g_i(x^*) = 0$, complementary slackness condition.

Proof: recall that $F(x) = \max\{f(x) - \mu, g_1(x), \dots, g_n(x)\}$. By the previous Lemma L15-b, $F(x^*) = 0 = \min F(\mathbb{R}^m)$. Hence, $0 \in \partial F(x^*) = \operatorname{conv}\left(\bigcup_{i \in A(x)} \partial g_i(x^*)\right)$ where $A(x^*) \coloneqq \{i \in \{0, 1, \dots, n\} : g_i(x^*) = 0 = F(x^*)\}$ by Fermat's theorem and Fact L15-c. Observe that $\partial g_0 = \partial f$ since $g_0 = f - \mu$. Hence $(\forall i \in A(x^*)) \exists \alpha_i \geq 0$ such that $\sum_{i \in A(x^*)} \alpha_i = 1$ and by convex combination,

$$\begin{aligned} 0 &\in \sum_{i \in A(x^*)} \alpha_i \, \partial g_i(x^*) \\ &= \alpha_0 \, \partial g_0(x^*) + \sum_{i \in A(x^*) \setminus \{0\}} \alpha_i \partial g_i(x^*) \\ &= \partial f(x^*) + \sum_{i \in A(x^*) \setminus \{0\}} \alpha_i \partial g_i(x^*) \end{aligned}$$

Now, for $i \in I \setminus A(x^*)$, set $\alpha_i = 0$. For $i \in I \setminus A(x^*)$, then $g_i(x^*) = 0$. And we can meet the complementary slackness condition. $\square$

Theorem L16-a, KKT conditions necessary part: Suppose $f, g_1, \dots, g_n$ are convex, x^* solves (P). Suppose that **Slater's condition** holds, i.e. $\exists$ slater point $s \in \mathbb{R}^m$ such that $(\forall i \in I = \{1, \dots, n\})\ g_i(s) < 0$. Then $\exists \lambda_1, \dots, \lambda_m \geq 0$ such that KKT conditions hold:

1. **Stationarity Condition**: $0 \in \partial f(x^*) + \sum_{i \in I} \lambda_i \, \partial g_i(x^*)$
2. **Complementary Slackness Condition**: $(\forall i \in I)\ \lambda_i g_i(x^*) = 0$

> Note that necessary part doens't necessarily require convexity, but the sufficient part does.

Proof: for part 1, recall Fritz-John, $\exists \alpha_0, \dots, \alpha_n \geq 0$, not all 0 such that $0 \in \alpha_0 \, \partial f(x^*) + \sum_{i \in I} \alpha_i \, \partial g_i(x^*) \quad (\mathrm{I.})$. It suffice to show that $\alpha_0 > 0$! Suppose for eventual contradiction that $\alpha_0 = 0$. By $(\mathrm{I.})$, $(\forall i \in I)\ \exists y_i \in \partial g_i(x^*)$ such that $\sum_{i \in I} \alpha_i y_i = 0$. Hence, $(\forall i \in I)(\forall y \in \mathbb{R}^m)\ g_i(x^*) + \langle y_i, y - x^* \rangle \leq g_i(y)$. In particular, $g_i(x^*) + \langle y_i, s - x^* \rangle \leq g_i(s)$. Multiplying both sides by $\alpha_i \geq 0$, and summing for all $i \in I$, and we have

$$0 = \sum_{i \in I} \alpha_i g_i(x^*) + \left\langle \sum_{i \in I} \alpha_i y_i, s - x^* \right\rangle \leq \sum_{i \in I} \alpha_i g_i(s) < 0$$

which is absurd. Hence $a_0 > 0$. Now divide both sides of $(\mathrm{I.})$ by α_0 and set $\lambda_i = \alpha_i / \alpha_0 \geq 0$ and we're done. $\square$

Theorem L16-b, KKT conditions sufficient part: Suppose $f, g_1, \dots, g_n$ are convex and $x^* \in \mathbb{R}^m$ satisfies:

1. **Primal Feasibility**: $(\forall i \in I)\ g_i(x^*) \leq 0$
2. **Dual Feasibility**: $(\forall i \in I)\ \lambda_i \geq 0$
3. **Stationarity**: $0 \in \partial f(x^*) + \sum_{i \in I} \lambda_i \, \partial g_i(x^*)$
4. **Complementary Slackness**: $(\forall i \in I)\ \lambda_i g_i(x^*) = 0$

Then x^* solves (P).

Proof: Define $h(x) \coloneqq f(x) + \sum_{i \in I} \lambda_i g_i(x)$, we have $h(x)$ is convex because non-negative linear combination of convex functions is convex. Therefore,

$$\begin{aligned} (\forall x \in \mathbb{R}^m)\ \partial h(x) &= \partial \left(f + \sum_{i \in I} \lambda_i g_i \right)(x) \\ &= \partial f(x) + \sum_{i \in I} \lambda_i \, \partial g_i(x) \quad \text{sum rule} \end{aligned}$$

Consequently, $0 \in \partial h(x^*) = \partial f(x^*) + \sum_{i \in I} \lambda_i \, \partial g_i(x^*)$. By Fermat's Theorem, x^* is a global minimizer of h. Now, let x be feasible for (P), then $f(x^*) = f(x^*) + \sum_{i \in I} \lambda_i g_i(x^*) = h(x^*) \leq h(x) = f(x) + \sum_{i \in I} \lambda_i g_i(x) \leq f(x)$. $\square$

Subgradient Method

Gradient Descent Classical Theory

Consider the problem $(\mathrm{P}) \quad \min\limits_{x\in\mathbb{R}^m} \quad f(x)$.

Descent Direction: Let $f:\mathbb{R}^m\to(-\infty,+\infty]$ be proper and let $x\in\operatorname{int}(\operatorname{dom}(f))$. $d\in\mathbb{R}^m\backslash\{0\}$ is a descent direction of f at x if the directional derivative satisfies $\nabla f(x)\cdot d<0$.

> Remarks L16-4:
>
> 1. If $0\neq\nabla f(x)$ exists at $x \implies -\nabla f(x)$ is a descent direction.
> 2. Let d be a descent direction, then $\exists\epsilon>0$ such that $(\forall 0<t\leq\epsilon)\ f(x+td)<f(x)$

Gradient/Steepest Descent Method: $x_0\in\mathbb{R}^m$. $(\forall n\in\mathbb{N})$ update via

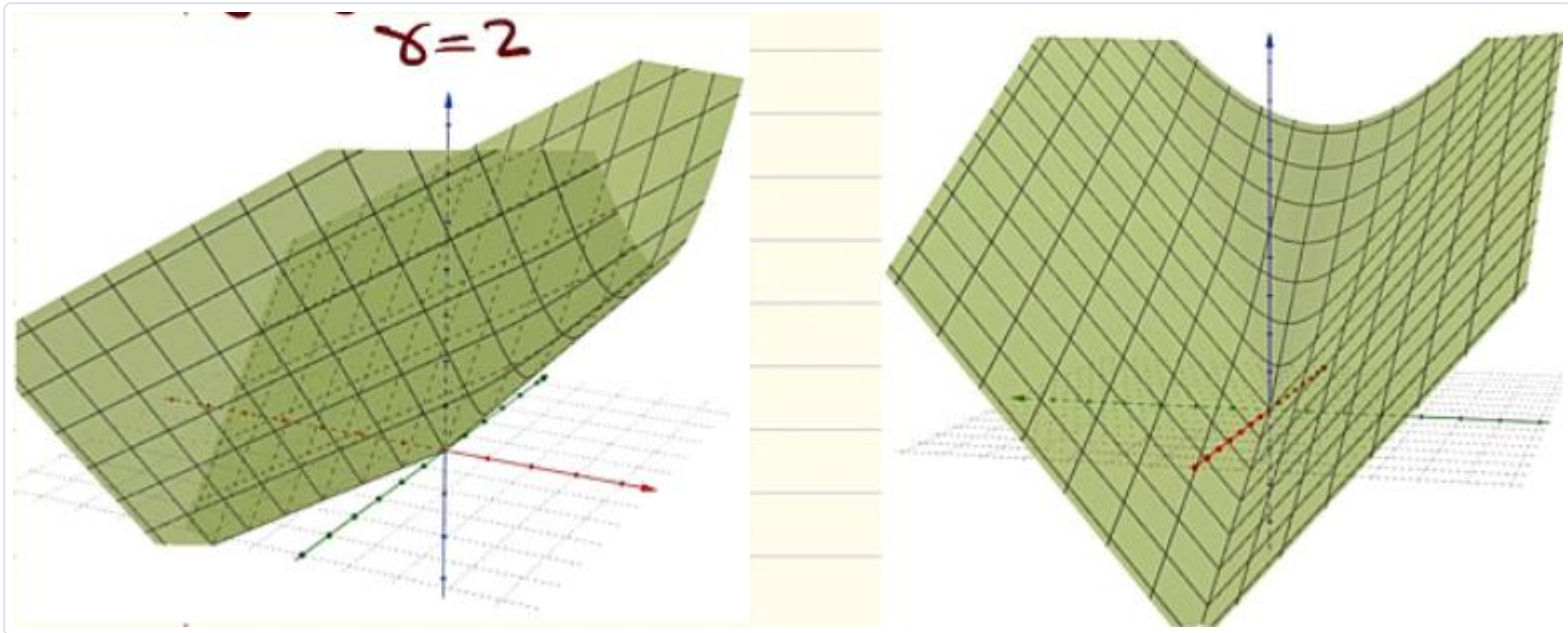
$$x_{n+1}:=x_n-t_n\nabla f(x_n)$$

where $t_n\in\operatorname*{arg\,min}\limits_{t\geq 0} f(x_n-t\nabla f(x_n))$. If f is strictly convex and coercive, then $x_n\to$ the unique minimizer (Peressin, Sullivan, Uhl).

Example L16-5, In the lack of smoothness: negative subgradients are not necessarily descent directions.

Consider $f:\mathbb{R}^2\to[0,+\infty):(x_1,x_2)\mapsto|x_1|+2|x_2|$, f is convex (sum of convex functions), and f is full domain (continuous). $\partial f(1,0)=\{1\}\times[-2,2]\ni(1,2)$. Let $d=(-1,-2)$ and let $t>0$, then $f((1,0)+t(-1,-2))=|1-t|+4|t|$. We see that $\nabla f(1,0)\cdot(-1,-2)=3>0$. Hence d is not a descent direction. □

Example L16-6(Wolfe): Let $\gamma>1$. Consider the function $f:\mathbb{R}^2\to\mathbb{R}:(x_1,x_2)\mapsto\begin{cases}\sqrt{x_1^2+\gamma x_2^2}, & |x_2|\leq x_1\\ \frac{x_1+\gamma|x_2|}{\sqrt{1+\gamma}}, & \text{otherwise}\end{cases}$. Observe that $\operatorname*{arg\,min}\limits_{x\in\mathbb{R}^m} f(x)=\emptyset$. Indeed, $\inf\limits_{x\in\mathbb{R}^m} f(x)=-\infty$ as $f(r,0)=\frac{r}{\sqrt{1+\gamma}}\to-\infty$ as $r\to-\infty$.



One can show that $f=\sigma_C$ where $C=\{(x_1,x_2)\in\mathbb{R}^2:x_1^2+\frac{x_2^2}{\gamma}\leq 1, x_1\geq\frac{1}{\sqrt{1+\gamma}}\}$. Therefore, f is convex. Also, f is differentiable on $D:=\mathbb{R}^2\backslash((-\infty,0]\times\{0\})$. Now, let $x_0=(\gamma,1)\in D$. The steepest descent will generate a sequence (details omitted) where $x_n=(\gamma(\frac{\gamma-1}{\gamma+1})^n,(-(\frac{\gamma-1}{\gamma+1}))^n)\to(0,0)$. Observe that $(0,0)$ is not a minimizer of f! In the absence of smoothness a lot of pathologies happen.

Projected Subgradient Method

Consider the problem $(\mathrm{P}) \quad \min(f(x):x\in C)$, where

- $f:\mathbb{R}^m\to(-\infty,+\infty]$ is convex, l.s.c. and proper.
- $C\neq\emptyset$ is convex, closed subset of $\operatorname{int}(\operatorname{dom}(f))$
- $S:=\operatorname*{arg\,min}\limits_{x\in C} f(x)\neq\emptyset$
- $\mu:=\min\limits_{x\in C} f(x)$
- $\exists L>0$ such that $(\forall c\in C)(\forall u\in\partial f(c))\ \|u\|\leq L$, i.e. $\sup\limits_{c\in C}\|\partial f(c)\|\leq L<+\infty$

Get $x_0\in C$. $(\forall n\in\mathbb{N})$, given x_n, pick a stepsize $t_n>0$, and let $f'(x_n)\in\partial f(x_n)$ (subgradient). We update via

$$x_{n+1}:=P_C(x_n-t_nf'(x_n))$$

> Since $C \subseteq \operatorname{int}(\operatorname{dom}(f))$, $(\forall n \in \mathbb{N})\ x_n \in \operatorname{int}(\operatorname{dom}(f))$. Therefore $\partial f(x_n) \neq \emptyset$, and $(x_n)_{n\in\mathbb{N}}$ is well-defined.

Lemma L17-a: Let $s \in S = \arg\min_{x\in C} f(x)$. Then

$$\|x_{n+1} - s\|^2 \leq \|x_n - s\|^2 - 2t_n(f(x_n) - \mu) + t_n^2\|f'(x_n)\|^2$$

Observe that $S \subseteq C$.

Proof:

$$\begin{aligned}
\|x_{n+1} - s\|^2 &= \|P_C(x_n - t_n f'(x_n)) - P_C(s)\|^2 \\
&\leq \|x_n - t_n f'(x_n) - s\|^2 \quad \because P_C \text{ is f.n.e.} \\
&= \|x_n - s\|^2 + t_n^2\|f'(x_n)\|^2 - 2t_n\langle x_n - s, f'(x_n)\rangle
\end{aligned}$$

So it remains to show that $-2t_n\langle x_n - s, f'(x_n)\rangle \leq -2t_n(f(x_n) - \mu)$, equivalently, $\langle x_n - s, f'(x_n)\rangle \geq f(x_n) - \mu$ which is true by the subgradient inequality $\mu = f(s) \geq f(x_n) + \langle f'(x_n), s - x_n\rangle$. □

What is a good step size t_n? Let us minimize the upper bound:

$$\begin{aligned}
0 &= \frac{d}{dt_n}\text{RHS of Lemma L17-a} \\
&= -2(f(x_n) - \mu) + 2t_n\|f'(x_n)\|^2
\end{aligned}$$

Assuming $f'(x_n) \neq 0$, (otherwise $0 \in \partial f(x_n)$, and by Fermat Theorem x_n is a global minimizer and we are done), solving the above for t_n and we get

$$t_n = \frac{f(x_n) - \mu}{\|f'(x_n)\|^2}$$

which is known as **Polyak's Rule**.

Theorem L17-b, consequence of projected subgradient method: we have

1. $(\forall s \in S)(\forall n \in \mathbb{N})\ \|x_{n+1} - s\| \leq \|x_n - s\|$, i.e. $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to S.
2. $f(x_n) \to \mu$
3. $\mu_n - \mu \leq \dfrac{L \cdot ds(x_0)}{\sqrt{n+1}} \to 0$ as $n \to 0$, where $\mu_n \coloneqq \min_{0\leq k\leq n} f(x_n)$
4. Let $\epsilon > 0$. If $n \geq \dfrac{L^2 d_S^2(x_0)}{\epsilon^2} - 1 \implies \mu_n \leq \mu + \epsilon$

Proof: For part 1, let $s \in S, n \in \mathbb{N}$

$$\begin{aligned}
\|x_{n+1} - s\|^2 &\leq \|x_n - s\|^2 - 2t_n(f(x_n) - \mu) + t_n^2\|f'(x_n)\|^2 \\
&= \|x_n - s\|^2 - 2\frac{(f(x_n) - \mu)^2}{\|f'(x_n)\|^2} + \frac{(f(x_n) - \mu)^2}{\|f'(x_n)\|^2} \\
&\leq \|x_n - s\|^2 - \frac{(f(x_n) - \mu)^2}{\|f'(x_n)\|^2} \\
&\leq \|x_n - s\|^2 - \frac{(f(x_n) - \mu)^2}{L^2} \\
&\leq \|x_n - s\|^2
\end{aligned}$$

For part 2, observe that $(\forall k \in \mathbb{N})\ \dfrac{(f(x_k) - \mu)^2}{L^2} \leq \|x_k - s\|^2 - \|x_{k+1} - s\|^2$. Summing the above inequalities over $k = 0$ to $k = n$ yields $\dfrac{1}{L^2}\displaystyle\sum_{k=0}^{n}(f(x_k) - \mu)^2 \leq \|x_0 - s\|^2 - \|x_{n+1} - s\|^2 \leq \|x_0 - s\|$. Letting $n \to \infty$, we learn that $0 \leq \displaystyle\sum_{k=0}^{\infty}(f(x_k) - \mu)^2 \leq L^2\|x_0 - s\|^2 < +\infty \implies f(x_k) - \mu \to 0$ by calculus, i.e. $f(x_k) \to \mu$.

For Part 3, recall $(\forall n \in \mathbb{N})\ \mu_n \coloneqq \min_{0\leq k\leq n} f(x_k)$. Let $n \geq 0$. Then $\forall k \in \{0, \ldots, n\}, (\mu_n - \mu)^2 \leq (f(x_k) - \mu)^2 \implies (n+1)\dfrac{(\mu_n - \mu)^2}{L^2} \leq \dfrac{1}{L^2}\displaystyle\sum_{k=0}^{n}(f(x_k) - \mu)^2 \leq \|x_0 - s\|^2$. Minimizing over $s \in S$, we get $(n+1)\dfrac{(\mu_n - \mu)^2}{L^2} \leq d_S^2(x_0)$.

For part 4, $n \geq \dfrac{L^2 d_S^2(x_0)}{\epsilon^2} - 1 \iff \dfrac{d_S^2(x_0)L^2}{(n+1)} \leq \epsilon^2$. And so from part 3 we get $(\mu_n - \mu)^2 \leq \dfrac{d_S^2(x_0)L^2}{(n+1)} \leq \epsilon^2$ which implies $\mu_n \leq \mu + \epsilon$.

Theorem L17-c, Convergence of Projected Subgradient: Suppose that $(x_n)_{n\in\mathbb{N}}$ is generated using projected subgradient method. Then $x_n \to$ a solution of (P) in S.

Proof: by the previous Theorem, $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to S. Since $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to S, $(x_n)_{n\in\mathbb{N}}$ is bounded. Also, by the previous theorem, $f(x_n) \to \mu = \min\limits_{x\in C} f(x)$. By the Bolzano-Weierstrass, $\exists x_{k_n} \to \bar{x}$, and $\bar{x} \in C$ because $(x_n)_{n\in\mathbb{N}}$ lies in C which is closed. Now $\mu = \min\limits_{x\in C} f(x) \leq f(\bar{x}) \leq \liminf\limits_{n\to\infty} f(x_{k_n}) = \mu$. Note that limit inferior comes from the definition of l.s.c. Then, $f(\bar{x}) = \mu$. Hence $\bar{x} \in S$. That is, all cluster points of $(x_n)_{n\in\mathbb{N}}$ lies in S. $x_n \to \bar{x} \in S$ by the Fejér monotone Theorem. □

Example L18-1: Let $C \subseteq \mathbb{R}^m$ be convex, closed and nonempty and let $x \in \mathbb{R}^m$. Then

$$\partial d_C(x) = \begin{cases} \frac{x - P_C(x)}{d_C(x)}, & x \notin C \\ N_C(x) \cap B(0,1), & x \in C \end{cases}$$

Consequently, $(\forall x \in \mathbb{R}^m)\ \sup \|\partial d_C(x)\| \leq 1$.

Lemma L18-a: Let f be convex, l.s.c., proper and let $\lambda > 0$. Then $\partial(\lambda f) = \lambda \partial f$.

The Convex Feasibility Problem

Given K closed convex subsets S_i of $\mathbb{R}^m$ such that $S = \bigcap\limits_{i=1}^{k} S_i \neq \emptyset$.

Problem: Find $x \in S$

Can we use the projected subgradient method? Recall the parameters from projected subgradient method, set $C = \mathbb{R}^m$, $P_C = \mathrm{Id}$. Set

$$f(x) = \max\{d_{S_1}(x), \ldots, d_{S_k}(x)\}$$

Then $(\forall x \in \mathbb{R}^m)\ f(x) \geq 0 \iff x \in S \neq \emptyset \implies \mu := \min_{x\in\mathbb{R}^m} f(x) = 0$. Additionally, $L = 1$ by the previous example.

Finally, observe that the max formula for subdifferentials implies that $x \notin S$.

Finally, we use the formula of ∂d_C for the case $x \notin S$ (since if $x \in S$ then we're done)

$$\begin{aligned} \partial f(x) &= \mathrm{conv}\{\partial d_{S_i}(x) : d_{S_i}(x) = f(x)\} \\ &= \mathrm{conv}\left\{ \frac{x - P_{S_i}(x)}{d_{S_i}(x)} : d_{S_i}(x) = f(x) \right\} \end{aligned}$$

What do we do with that? Well, given point x_n, we pick an index i_n such that $d_{S_{i_n}}(x_n) = f(x_n)$. Set $f'(x_n) := \dfrac{x_n - P_{S_{i_n}}(x_n)}{d_{S_{i_n}}(x_n)}$.

What about t_n? Polyak's step size:

$$t_n = \frac{f(x_n) - \mu}{\|f'(x_n)\|^2} = d_{S_{i_n}}(x_n)$$

The update leads to the **Greedy Projection Algorithm**:

$$\begin{aligned} x_{n+1} &= x_n - t_n\ f'(x_n) \\ &= P_{S_{i_n}}(x_n) \end{aligned}$$

where S_{i_n} is any set that is farthest away from x_n. And by Theorem L17-c, (Convergence of Projected Subgradient), $x_n \to$ some point in S.

> In the case $K = 2$, we have **the method of alternating projections (MAP)**. $x_0 \in \mathbb{R}^m$. Update via $x_{n+1} = P_{S_2} \circ P_{S_1}(x_n)$

Example L18-3: Find $x \in S$ where $S := \{x \in \mathbb{R}^m : Ax = b, x \geq 0\}$, where A is $k \times m$ matrix and $b \in \mathbb{R}^k$. We can use MAP! Set $S_1 = \mathbb{R}^m_+, P_{S_1}(x) = x^+ = (\max\{\zeta_i, 0\})_{i=1}^m$. We also have $S_2 = \{x \in \mathbb{R}^m : Ax = b\} = A^{-1}(b)$ (note that this is not the matrix inverse, but the **inverse image** of b). $P_{S_2}(x) = x - A^\dagger(Ax - b)$.

$A^\dagger$ is the **Moore-Penrose pseudo inverse** (pinv). From Wikipedia. Let $x_0 \in \mathbb{R}^m$. Update via $x_{n+1} = P_{S_2} \circ P_{S_1}(x_n) = P_{S_2}(x_n^+) = x_n^+ - A^\dagger(Ax_n^+ - b) \to \bar{x} \in S$.

> Overall, MAP is rather easy to visualize.

Remark L18-4: In practice, it is possible that $\mu \coloneqq \min_{x\in C} f(x)$ is NOT known to us. In this case replace Polyak's stepsize by a sequence $(t_n)_{n\in\mathbb{N}}$ such that

$$\frac{\sum_{k=0}^{n} t_k^2}{\sum_{k=0}^{n} t_k} \to 0 \quad \text{as} \quad n \to +\infty$$

For example, $t_k = \frac{1}{k+1}$. One can show that $\mu_n \coloneqq \min\{f(x_0), \dots, f(x_n)\} \to \mu$ as $n \to \infty$.

Proximal Gradient Method (PGM)

Consider the problem

$$(\mathrm{P}) \quad \min_{x\in\mathbb{R}^m} F(x) \coloneqq f(x) + g(x)$$

Assumptions: (P) has solutions $s \coloneqq \operatorname{arg\,min}_{x\in\mathbb{R}^m} F(x) \neq \emptyset$. $\mu = \min_{x\in\mathbb{R}^m} F(x)$.

- f is "nice": convex, l.s.c., proper, differentiable on $\operatorname{int}(\operatorname{dom}(f)) \neq \emptyset$, ∇f is L-Lipschitz.
- g is convex, l.s.c., proper, $\operatorname{dom}(g) \subseteq \operatorname{int}(\operatorname{dom}(f))$

> Note that $\operatorname{ri}(\operatorname{dom}(g)) \subseteq \operatorname{dom}(g) \subseteq \operatorname{ri}(\operatorname{dom}(f)) \implies \operatorname{ri}(\operatorname{dom}(g)) \cap \operatorname{ri}(\operatorname{dom}(f)) = \operatorname{ri}(\operatorname{dom}(g)) \neq \emptyset$. This means we can use the subdifferential sum rule.

Example L18-5: $\min_{x\in C} f(x)$, $\emptyset \neq C \subseteq \mathbb{R}^m$ is convex and closed. Recall that this is equivalent to $\min_{x\in\mathbb{R}^m} f(x) + \delta_C(x)$.

PGM: $x \in \operatorname{int}(\operatorname{dom}(f)) \supseteq \operatorname{dom}(g)$. Update via

$$\begin{aligned} x_{n+1} &= \operatorname{prox}_{\frac{1}{L}g}\left(x_n - \frac{1}{L}\nabla f(x_n)\right) \\ &= \operatorname*{arg\,min}_{y\in\mathbb{R}^m}\left\{\frac{1}{L}g(y) + \frac{1}{2}\left\|y - \left(x_n - \frac{1}{L}\nabla f(x_n)\right)\right\|\right\} \\ &\in \operatorname{dom}(g) \subseteq \operatorname{int}(\operatorname{dom}(f)) = \operatorname{dom}(\nabla f) \end{aligned}$$

Set $T = \operatorname{prox}_{\frac{1}{L}g} \circ \left(\mathrm{Id} - \frac{1}{L}\nabla f\right)$, i.e. $(\forall x \in \mathbb{R}^m)\ Tx = \operatorname{prox}_{\frac{1}{L}g}(x - \frac{1}{L}\nabla f(x))$.

Theorem L18-b: let $x \in \mathbb{R}^m$. Then $x \in S = \operatorname*{arg\,min}_{x\in\mathbb{R}^m}(F) = \operatorname*{arg\,min}_{x\in\mathbb{R}^m}(f+g) \iff x = Tx$ (i.e. $x \in \operatorname{Fix}(T)$)

Proof: Observe that by Fermat,

$$\begin{aligned} & x \in S \\ \iff & 0 \in \partial(f+g)(x) = \partial f(x) + \partial g(x) = \nabla f(x) + \partial g(x) \\ \iff & -\nabla f(x) \in \partial g(x) \\ \iff & x - \frac{1}{L}\nabla f(x) \in x + \partial\left(\frac{1}{L}g\right)(x) = \left(\mathrm{Id} + \partial\left(\frac{1}{L}g\right)\right)(x) \\ \iff & x \in \left(\mathrm{Id} + \partial\left(\frac{1}{L}g\right)\right)^{-1}\left(x - \frac{1}{L}\nabla f(x)\right) \\ \iff & x = \operatorname{prox}_{\frac{1}{L}g}\left(\mathrm{Id} - \frac{1}{L}\nabla f\right)(x) = Tx \end{aligned}$$

where T is as defined earlier. □

Fact L18-c: Let $f : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper and let $\beta > 0$. Then f is β-strongly convex $\iff (\forall x \in \operatorname{dom}(\partial f))(\forall v \in \partial f(x))\ f(y) \geq f(x) + \langle v, y - x\rangle + \frac{\beta}{2}\|y - x\|^2$.

The Prox-Grad Inequality

Proposition L18-d: Let $x \in \mathbb{R}^m, y \in \operatorname{int}(\operatorname{dom}(f)), y_o \coloneqq Ty = \operatorname{prox}_{\frac{1}{L}g}(y - \nabla f(y))$. Then

$$F(x) - F(y_o) \geq \frac{L}{2}\|x - y_o\|^2 - \frac{L}{2}\|x - y\| + D_f(x, y)$$

where the **Bregman distance**, $D_f(x, y) \coloneqq f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geq 0$ holds by convexity of f.

Proof: Define $h(z) \coloneqq f(y) + \langle \nabla f(y), z - y \rangle + g(z) + \frac{L}{2}\|z-y\|^2$. Then h is L-strongly convex (sum of convex and strongly convex). Let $z \in \mathbb{R}^m$. Then

$$
\begin{aligned}
& z \text{ minimizes } h \\
\iff & 0 \in \partial(f(y) + \langle \nabla f(y), z-y\rangle + g(z) + \frac{L}{2}\|z-y\|^2) \\
& = \nabla f(y) + \partial g(z) + L(z-y) \\
\iff & y - \frac{1}{L}\nabla f(y) \in z + \partial\left(\frac{1}{L}g\right)(z) = \left(\mathrm{Id} + \partial\left(\frac{1}{L}g\right)\right)(z) \\
\iff & z = Ty =: y_o \implies \arg\min h = \{y_o\}
\end{aligned}
$$

Recall Fact L18-c, we have $h(x) - h(y_o) \geq \frac{L}{2}\|x-y_o\|^2 \quad (\mathrm{I.})$. Moreover, by the descent lemma we have $f(y_o) \leq f(y) + \langle \nabla f(y), y_o - y\rangle + \frac{L}{2}\|y_o - y\|^2$. Therefore, $h(y_o) = f(y) + \langle \nabla f(y), y_o - y\rangle + g(y_o) + \frac{L}{2}\|y_o-y\|^2 \geq f(y_o) + g(y_o) = F(y_o)$. Combining with $(\mathrm{I.}), h(x) - F(y_o) \geq h(x) - h(y_o) \geq \frac{L}{2}\|x-y_o\|^2$. Using the definition of h and add $f(x)$ to both sides yield the desired result. □

Lemma 19-a, Sufficient Decrease Lemma: using the same convention as Proposition 18-d, we have

$$F(y_o) \leq F(y) - \frac{L}{2}\|y - y_o\|^2$$

Proof: use Proposition 18-d with x replaced by y and recall that because f is convex, $D_f(x,y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle \geq 0$. □

The **proximal gradient method**: Given $y \in \mathrm{int}(\mathrm{dom}(f))$, update via $y_o \coloneqq \mathrm{Prox}_{\frac{1}{L}g}(y - \frac{1}{L}\nabla f(y)) =: Ty \in \mathrm{dom}(g) \subseteq \mathrm{int}(\mathrm{dom}(f)) = \mathrm{dom}(\nabla f)$.

The algorithm: given $x_0 \in \mathrm{int}(\mathrm{dom}(f))$. $(\forall n \in \mathbb{N})$. Update via $x_{n+1} \coloneqq Tx_n = \mathrm{Prox}_{\frac{1}{L}g}(x_n - \frac{1}{L}\nabla f(x_n))$.

Theorem L19-b, $O(\frac{1}{n})$ rate of convergence of function values: the following holds:

1. $(\forall s \in S)(\forall n \in \mathbb{N})\ \|x_{n+1} - s\| \leq \|x_n - s\|$, i.e. $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone w.r.t. S.
2. $(F(x_n))_{n\in\mathbb{N}}$ decreases to μ, more precisely, $0 \leq F(x_n) - \mu \leq \frac{L \cdot d_S^2(x_0)}{2n} = O(\frac{1}{n})$. Hence $F(x_n) \to \mu$.

Proof: applying Lemma 19-1 yields $F(x_{n+1}) \leq F(x_n) - \frac{L}{2}\|x_{n+1} - x_n\|^2 \leq F(x_n)$.

(i.) Recalling let $s \in S$, let $k \in \mathbb{N}$.Applying Proposition L18-d yields $0 \geq \underbrace{F(s)}_{=\mu} - F(x_{k+1}) \geq \frac{L}{2}\|s - x_{k+1}\|^2 - \frac{L}{2}\|s - x_k\|^2 \implies$ $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone w.r.t. S.

(ii.) Multiplying by $\frac{2}{L}$ and adding the resulting inequalities from $k = 0$ to $k = n-1$ and telescoping yields $\frac{2}{L}\left(\sum_{k=0}^{n-1}(\mu - F(x_{k+1}))\right) \geq \|s - x_n\|^2 - \|s - x_0\|^2 \geq -\|s - x_0\|^2$. In particular, setting $s = P_S(x_0) \in S$.

We obtain $d_S^2(x_0) = \|P_S(x_0) - x_0\|^2 \geq \frac{2}{L}\sum_{k=0}^{n-1}(F(x_{k+1}) - \mu) \geq \frac{2}{L}\sum_{k=0}^{n-1}(F(x_n) - \mu) = \frac{2}{L}n(F(x_n) - \mu)$.

Equivalently, $0 \leq F(x_n) - \mu \leq \frac{L d_S^2(x_0)}{2n}$, and $F(x_n) \to \mu$. □

Theorem 19-c, convergence of PGM: x_n converges to some solution in $S = \arg\min_{x\in\mathbb{R}^m} F(x)$.

Proof: By the previous theorem, we have $(x_n)_{n\in\mathbb{N}}$ is Fejér monotone w.r.t. S. Done if we can show that every cluster point of $(x_n)_{n\in\mathbb{N}}$ lies in S. Suppose that $\bar{x}$ is a cluster point of $(x_n)_{n\in\mathbb{N}}$, say $x_{k_n} \to \bar{x}$.

The goal is $F(\bar{x}) = \mu$. Indeed, $\mu \leq F(\bar{x}) \leq \liminf_{n\to\infty} F(x_{k_n}) = \mu \implies F(\bar{x}) = \mu \iff \bar{x} \in S$. Note that the $\liminf$ comes from the fact F is l.s.c. □

Proposition L19-d: the following holds:

1. $\frac{1}{L}\nabla f$ is f.n.e.
2. $\mathrm{Id} - \frac{1}{L}\nabla f$ is f.n.e.
3. $T = \mathrm{Prox}_{\frac{1}{L}g}(\mathrm{Id} - \nabla f)$ is 2/3-averaged.

Proof: for part 1. and 2., recalling Theorem L9-d, the Lipschitz continuous gradient characterization, dividing both sides of part 4 by $1/L$ yields $\langle \frac{1}{L}\nabla f(x) - \frac{1}{L}\nabla f(y), x - y\rangle \geq \|\frac{1}{L}\nabla f(x) - \frac{1}{L}\nabla f(y)\|^2$. From a result from assignment, this implies $\frac{1}{L}\nabla f$ is f.n.e.

> Note this fact: if f is convex, then ∇f is f.n.e.

For part 3., recall that $\mathrm{Prox}_{\frac{1}{L}g}$ is f.n.e. Hence, $\mathrm{Prox}_{\frac{1}{L}g}$ and $\mathrm{Id} - \frac{1}{L}\nabla f$ are both $\frac{1}{2}$-averaged. Consequently, the composition $\mathrm{Prox}_{\frac{1}{L}g} \circ (\mathrm{Id} - \frac{1}{L}\nabla f)$. (This again follows from an assignment result.)

Remark L19-e: Recalling Proposition L14-d, part 1 and 4, one can show that for $T = \mathrm{Prox}_{\frac{1}{L}g}(\mathrm{Id} - \frac{1}{L}\nabla f)$, we have $(\forall x \in \mathbb{R}^m)(\forall y \in \mathbb{R}^m)\ \frac{1}{2}\|(\mathrm{Id} - T)x - (\mathrm{Id} - T)y\|^2 \leq \|x - y\|^2 - \|Tx - Ty\|^2$.

Theorem L19-f: Recalling the PGM iteration, we have $\|x_{n+1} - x_n\| \leq \frac{\sqrt{2}d_S(x_0)}{\sqrt{n}} = O(\frac{1}{\sqrt{n}})$.

Proof: let $s \in S$, we observe taht $s = T_s$ by Theorem L18-b. By Remark L19-e with $x = x_k, y = s \in S$ we get $\frac{1}{2}\|(\mathrm{Id} - T)x_k - \underbrace{(\mathrm{Id} - T)s}_{=0}\| \leq \|x_k - s\|^2 - \|\underbrace{Tx_k}_{x_{k+1}} - \underbrace{Ts}_{=s}\|^2$. That is, $\frac{1}{2}\|x_k - x_{k+1}\|^2 \leq \|x_k - s\|^2 - \|x_{k+1} - s\|^2 \quad (\mathrm{I.})$.

Using the previous proposition T is 2/3-averaged, hence T is nonexpansive. Therefore: $\|x_k - x_{k+1}\| \leq \|x_{k-1} - x_k\| \leq \cdots \leq \|x_0 - x_1\|$. Summing $(\mathrm{I.})$ over $k = 0$ to $k = n - 1$ we have $\|x_0 - s\|^2 - \|x_n - s\|^2 \geq \frac{1}{2}\sum_{k=0}^{n-1}\|x_k - x_{k+1}\|^2 \geq \frac{1}{2}n \cdot \|x_{n-1} - x_n\|^2$.

In particular, for $s = P_S(x_0)$, we get $\frac{1}{2}n\|x_{n-1} - x_n\|^2 \leq d_S^2(x_0) \implies \|x_{n-1} - x_n\| \leq \frac{\sqrt{2}}{\sqrt{n}}d_S(x_0) = O(\frac{1}{\sqrt{n}})$. □

Corollary L19-7, the classical proximal point algorithm, (1970's Rockafeller): $g : \mathbb{R}^m \to (-\infty, +\infty]$ is convex, l.s.c. and proper, $c > 0$.

$$(\mathrm{P}) \quad \min_{x \in \mathbb{R}^m} g(x)$$

Assume that $s \coloneqq \arg\min_{x \in \mathbb{R}^m} g(x) \neq \emptyset$. Let $x_o \in \mathbb{R}^m$. Update via $x_{n+1} = \mathrm{Prox}_{c \cdot g} x_n$. Then

- $g(x_n) \to \mu^+ = \min g(\mathbb{R}^m)$
- $0 \leq g(x_n) - \mu \leq \frac{d_S^2(x_0)}{2cn}$
- $x_n \to$ some point in S
- $\|x_{n-1} - x_n\| \leq \frac{\sqrt{2}d_S(x_0)}{\sqrt{n}}$

Proof: set $(\forall x \in \mathbb{R}^m)\ f(x) = 0$. Then $(\forall x \in \mathbb{R}^m)\ \nabla f(x) = 0 \implies \nabla f$ is L-Lipschitz for any $L > 0$. In particular, for $L = \frac{1}{c} > 0$.

Observe that P can be written as $\min_{x \in \mathbb{R}^m} \underbrace{f(x) + g(x)}_{F(x) = g(x)} \implies S = \arg\min_{x \in \mathbb{R}^m} F(x) = \arg\min_{x \in \mathbb{R}^m} g(x)$. Additionally, $\nabla f = 0 \implies$ $\mathrm{Id} - \frac{1}{L}\nabla f = \mathrm{Id} \implies T = \mathrm{Prox}_{\frac{1}{L}g}(\mathrm{Id} - \frac{1}{L}\nabla f) = \mathrm{Prox}_{cg} \circ \mathrm{Id} = \mathrm{Prox}_{cg}$. And we are done by previous results. □

Fast Iterative Shrinkage Thresholding Algorithm (FISTA)

Consider the problem $(\mathrm{P}) \quad F(x) \coloneqq f(x) + g(x)$ with the following assumptions:

- (P) has solutions, and $S \coloneqq \arg\min_{x \in \mathbb{R}^m} F(x) \neq \emptyset$, $\mu = \min_{x \in \mathbb{R}^m} F(x)$
- f is "nice": convex, l.s.c., proper, and differentiable on $\mathbb{R}^m$; ∇f is L-Lipschitz on $\mathbb{R}^m$
- g is convex, l.s.c. and proper.

FISTA: Let $x_0 \in \mathbb{R}^m, t_0 = 1, y_0 = x_0$. Update via

$$\begin{aligned}
t_{n+1} &= \frac{1 + \sqrt{1 + 4t_n^2}}{2} \quad \text{Note that } t_{n+1}^2 - t_{n+1} = t_n^2 \\
x_{n+1} &= \mathrm{Prox}_{\frac{1}{L}g} \circ \left(\mathrm{Id} - \frac{1}{L}\nabla f\right)(y_n) \\
y_{n+1} &= x_{n+1} + \frac{t_n - 1}{t_{n+1}}(x_{n+1} - x_n) \\
&= \left(1 - \frac{1 - t_n}{t_{n+1}}\right)x_{n+1} + \frac{1 - t_n}{t_{n+1}}x_n \\
&\in \mathrm{aff}\{x_n, x_{n+1}\}
\end{aligned}$$

Remark L20-a: The sequence $(t_n)_{n \in \mathbb{N}}$ satisfies $(\forall n \in \mathbb{N})\ t_n \geq \frac{n+2}{2} \geq 1$. This follows from induction.

Theorem L20-b, $O(1/n^2)$ convergence rate for FISTA: $0 \leq F(x_n) - \mu \leq \frac{2Ld_S^2(x_0)}{(n+1)^2} = O(1/n^2)$.

> Note that this has a faster convergence rate than PGM.

Proof: set $s = P_S(x_0)$, by convexity of F we have $F(\frac{1}{t_n}s + (1 - \frac{1}{t_n})x_n) \leq \frac{1}{t_n}F(s) + (1 - \frac{1}{t_n})F(x_n)$. Set $(\forall n \in \mathbb{N})\ \delta_n = F(x_n) - \mu \geq 0$. Observe that

$$
\begin{aligned}
&\left(1-\frac{1}{t_n}\right)\delta_n-\delta_{n+1}\\
=&\left(1-\frac{1}{t_n}\right)F(x_n)+\frac{1}{t_n}F(s)-F(x_{n+1})\\
\geq& F\left(\frac{1}{t_n}s+\left(1-\frac{1}{t_n}\right)x_n\right)-F(x_{n+1})\quad\text{(I.)}
\end{aligned}
$$

Apply Proposition L18-d, and using the definition of FISTA yields

$$
\begin{aligned}
&F\left(\frac{1}{t_n}s+\left(1-\frac{1}{t_n}\right)x_n\right)-F(x_{n+1})\\
\geq&\frac{L}{2}\left\|\frac{1}{t_n}s+\left(1-\frac{1}{t_n}\right)x_n-x_{n+1}\right\|^2-\frac{L}{2}\left\|\frac{1}{t_n}s+\left(1-\frac{1}{t_n}\right)x_n-y_n\right\|^2\\
=&\frac{L}{2t_n^2}\left\|t_nx_{n+1}-(s+(t_n-1)x_n)\right\|^2-\frac{L}{2t_n^2}\left\|t_ny_n-(s+(t_n-1)x_n)\right\|^2\quad\text{(II.)}
\end{aligned}
$$

$$
\begin{aligned}
&\|t_ny_n-(s+(t_n-1)x_n)\|^2\quad\text{sub. def. of }y_n\\
=&\|t_{n-1}x_n-(s+(t_{n-1}-1)x_{n-1})\|^2\quad\text{(III.)}
\end{aligned}
$$

Therefore, using the fact $t_{n+1}^2-t_{n+1}=t_n^2$, we learn that

$$
\begin{aligned}
&t_{n-1}^2\delta_n-t_n^2\delta_{n+1}\\
=&(t_n^2-t_n)\delta_n-t_n^2\delta_{n+1}\\
=&t_n^2\left(\left(1-\frac{1}{t_n}\right)\delta_n-\delta_{n+1}\right)\\
\geq&t_n^2\left(F\left(\frac{1}{t_n}s+\left(1-\frac{1}{t_n}\right)x_n\right)-F(x_{n+1})\right)\\
\geq&\frac{L}{2}\|t_nx_{n+1}-(s+(t_n-1)x_n)\|^2-\frac{L}{2}\|t_ny_n-(s+(t_n-1)x_n)\|^2\\
=&\frac{L}{2}\|\underbrace{t_nx_{n+1}-(s+(t_n-1)x_n)}_{=:u_{n+1}}\|^2-\frac{L}{2}\|\underbrace{t_{n-1}x_n-(s+(t_{n-1}-1)x_{n-1})}_{=:u_n}\|^2
\end{aligned}
$$

Multiplying both sides of the above inequality by $\frac{2}{L}$ and rearranging yield $\|u_{n+1}\|^2+\frac{2}{L}t_n^2\delta_{n+1}\leq\|u_n\|^2+\frac{2}{L}t_{n-1}^2\delta_n$, therefore

$$
\begin{aligned}
\frac{2}{L}t_n^2\delta_{n+1}&\leq\|u_n\|^2+\frac{2}{L}t_{n-1}^2\delta_n\\
&\leq\cdots\\
&\leq\|u_1\|^2+\frac{2}{L}t_0^2\delta_1\\
&=\|\underbrace{t_0}_{=1}x_1-(s+(t_0-1)x_0)\|^2+\frac{2}{L}(1)(F(x_1)-\mu)\\
&=\|x_1-s\|^2+\frac{2}{L}(F(x_1)-\mu)\\
&\leq\|x_0-s\|^2\quad\text{follows from Prop L18-d}
\end{aligned}
$$

Finally, by definition of δ_n, $F(x)-\mu=\delta_n\leq\frac{L}{2}\|x_0-s\|^2\cdot\frac{1}{t_{n-1}^2}\leq\frac{2Ld_S^2(x_0)}{(n+1)^2}$, using the fact $t_n\geq\frac{n+2}{2}$. □

Iterative Shrinkage Thresholding Algorithm (ISTA)

Special case of PGM with $g(x)=\lambda\|x\|_1,\lambda>0\implies\frac{1}{L}g(x)=\frac{\lambda}{L}\|x\|_1$, $\operatorname{Prox}_{\frac{1}{L}g}(x)=(\operatorname{Prox}_{\frac{\lambda}{L}\|\cdot\|_1}(x))_{i=1}^n=(\operatorname{sgn}(x_i)\cdot\max\{0,|x_i|-\frac{\lambda}{L}\})_{i=1}^n$

> FISTA is the accelerated version of ISTA.

Example L20-3, l_1 regularized least squares, consider the problem: $(\mathrm{P})\quad\min_{x\in\mathbb{R}^m}(\frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1,\lambda>0)$, A is $n\times m$ matrix.

- $g(x)=\lambda\|x\|_1$ is convex, l.s.c. and proper
- $f(x)=\frac{1}{2}\|Ax-b\|_2^2$ is smooth, $(\forall x\in\mathbb{R}^m)\quad\nabla f(x)=A^\intercal(Ax-b)$
- $\operatorname{dom}(f)=\operatorname{dom}(g)=\mathbb{R}^m$, ∇f is L-Lipschitz continuous $\iff\lambda_{\max}(\nabla^2f(x))\leq L\iff\lambda_{\max}(A^\intercal A)\leq L$. Take $L:=\lambda_{\max}(A^\intercal A)$ and we get the Lipschitz continuity.

- $S \neq \emptyset$: Indeed, $F(x) = f(x) + g(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1$ is continuous, convex, coercive, and $\mathrm{dom}(F) = \mathbb{R}^m$. This implies $S = \arg\min F \neq \emptyset$

> Computational tip: sometimes m is large and computing the eigenvalues of $A^{\itnercal}A$ ($m \times m$ matrix) is not so easy. In this case, you could use an upper bound on eigen values, e.g. the **Frobenius norm**:
>
> $$\|A\|_F^2 = \sum_{j=1}^{m}\sum_{i=1}^{n} a_{ij}^2 = \mathrm{trace}(A^\intercal A) = \sum_{i=1}^{m} \lambda_i(A^\intercal A)$$

Douglas-Rachford (DR) Operator

Consider the problem

$$(\mathrm{P}) \quad \min_{x\in\mathbb{R}^m} \underbrace{f(x) + g(x)}_{F(x)}$$

- f is convex, l.s.c., and proper
- g is convex, l.s.c., and proper
- $S = \arg\min_{x\in\mathbb{R}^m} F(x) \neq \emptyset$, no further assumptions of smoothness or domain inclusions.

Suppose that $\exists s \in S$ such that $0 \in \partial f(s) + \partial g(s) \subseteq \partial(f+g)(s)$. This always holds regardless whether prerequisite for sum rule is met, i.e. $\mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(\mathrm{dom}(g)) \neq \emptyset \implies \partial(f+g) = \partial f + \partial g$.

> Note that these requirements are much less strict than the projected subgradient method or PGM.

Recall in Asignment 4 that $\mathrm{Prox}_f = (\mathrm{Id} + \partial f)^{-1}, \mathrm{Prox}_g = (\mathrm{Id} + \partial g)^{-1}$. Set

$$\begin{aligned} R_f &= 2\mathrm{Prox}_f - \mathrm{Id} \\ R_g &= 2\mathrm{Prox}_g - \mathrm{Id} \end{aligned}$$

Define the Douglas-Rachford (DR) operator as follows:

$$\begin{aligned} T &= \mathrm{Id} - \mathrm{Prox}_f + \mathrm{Prox}_g(2\mathrm{Prox}_f - \mathrm{Id}) \\ &= \mathrm{Id} - \mathrm{Prox}_f + \mathrm{Prox}_g R_f \end{aligned}$$

Lemma L22-a, the following hold:

1. R_f, R_g is nonexpansive.
2. $T = \frac{1}{2}(\mathrm{Id} + R_g R_f)$
3. T is firmly nonexpansive (f.n.e.)

Proof: for part 1, recall that Prox_f is f.n.e. by Proposition L14-b. Now combine this with Assignment 3, Problem 3, we get the result.

For part 2, indeed

$$\begin{aligned} \frac{1}{2}(\mathrm{Id} + R_g R_f) &= \frac{1}{2}(\mathrm{Id} + (2\mathrm{Prox}_g - \mathrm{Id})R_f) \\ &= \frac{1}{2}(\mathrm{Id} + 2\mathrm{Prox}_g R_f - R_f) \\ &= \frac{1}{2}(\mathrm{Id} + 2\mathrm{Prox}_g R_f - (2\mathrm{Prox}_f - \mathrm{Id})) \\ &= \mathrm{Id} - \mathrm{Prox}_f + \mathrm{Prox}_g R_f =: T \end{aligned}$$

For part 3, observe that $R_g R_f = R_g \circ R_f$ is a composition of two nonexpansive mappings, hence $R_g R_f$ is nonexpansive. Therefore, $T = \frac{1}{2}\mathrm{Id} + \frac{1}{2}R_g R_f$. That is, T is $\frac{1}{2}$-averaged, equivalently, T is f.n.e. by Remark L12-7. □

Remark L22-b: $\mathrm{Fix}(T) = \mathrm{Fix}(R_g R_f)$.

Proposition L22-c: $\mathrm{Prox}_f(\mathrm{Fix}(T)) \subseteq S$.

Proof: Let $x \in \mathbb{R}^m$, and set $s = \mathrm{Prox}_f(x)$. On the one hand,

$$
\begin{aligned}
& s = \mathrm{Prox}_f(x) \\
\iff & x \in (\mathrm{Id} + \partial f)(s) = s + \partial f(s) \\
\iff & \underbrace{2\mathrm{Prox}_f(x)}_{=s} - \underbrace{(2\mathrm{Prox}_f(x) - x)}_{=R_f(x)} \in s + \partial f(s) \\
\iff & 2s - R_f(x) \in s + \partial f(s) \\
\iff & s - R_f(x) \in \partial f(s) \quad (\mathrm{I.})
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
& x \in \mathrm{Fix}(T) \\
\iff & x = Tx \\
\iff & x = x - \mathrm{Prox}_f(x) + \mathrm{Prox}_g \circ R_f(x) \\
\iff & \mathrm{Prox}_f(x) = \mathrm{Prox}_g \circ R_f(x) \\
\iff & s = \mathrm{Prox}_g \circ R_f(x) \\
\iff & R_f(x) \in s + \partial g(s), \quad \because \mathrm{Prox}_g = (\mathrm{Id} + \partial g)^{-1} \\
\iff & 0 \in s - R_f(x) + \partial g(s) \\
\iff & R_f(x) - s \in \partial g(s) \quad (\mathrm{II.})
\end{aligned}
$$

Altogether, the inclusions $(\mathrm{I.})$ and $(\mathrm{II.})$ imply that $0 \in \partial f(s) + \partial g(s) \subseteq \partial(f+g)(s) \implies s \in S = \underset{x \in \mathbb{R}^m}{\arg\min} F(x)$.

Theorem L22-d: Let $x_0 \in \mathbb{R}^m$, update via

$$
x_{n+1} \coloneqq x_n - \mathrm{Prox}_f(x_n) + \mathrm{Prox}_g(2\mathrm{Prox}_f(x_n) - x_n)
$$

Then $\mathrm{Prox}_f(x_n) \to$ a minimizer of $f + g$.

Proof: rewrite x_{n+1} as $x_{n+1} = T(x_n) = T^{n+1}(x_0)$. Then by Corollary L14-a, $x_{n+1} \to \bar{x} \in \mathrm{Fix}(T)$. Observe that Prox_f is (firmly) nonexpansive by Proposition L14-b, hence continuous by Proposition L12-d. Consequently, $\mathrm{Prox}_f(x_n)$ will converge to $\mathrm{Prox}_f(\bar{x}) =: s$. Finally, observe that $s \in \mathrm{Prox}_f(\mathrm{Fix}(T)) \subseteq S$ by proposition L22-c. $\square$

Stochastic Projected Gradient Method

Consider the problem $\min\limits_{x \in C} f(x)$ with assumptions:

- f is convex, l.s.c., and proper
- $\emptyset \neq C$ closed and convex, and $C \subseteq \mathrm{int}(\mathrm{dom}(f))$
- $S \coloneqq \underset{x \in C}{\arg\min} f(x) \neq \emptyset$

Set $\mu \coloneqq \min f(C)$, stochastic projected subgradient method states that given $x_0 \in C$, update via

$$
x_{n+1} \coloneqq P_C(x_n - t_n g_n)
$$

Assumption on t_n: $t_n > 0$ such that $\sum\limits_{n=0}^{\infty} t_n \to +\infty$ and $\dfrac{\sum\limits_{k=0}^{n} t_k^2}{\sum\limits_{k=0}^{n} t_k} \to 0$ as $k \to +\infty$. For example, $t_n = \dfrac{\alpha}{n+1}, \alpha > 0$.

What about g_n? Choose g_n to be a random vector such that the following assumptions are satisfied.

- **Unbiased Subgradient**: $(\forall n \in \mathbb{N})\ E(g_n | x_n) \in \partial f(x_n)$, where $E(\dots)$ denotes expectation (from probability) that g_n is a subgradient given x_n. Equivalently, $(\forall y \in \mathbb{R}^m)\ f(x_n) + \langle E(g_n|x_n), y - x_n \rangle \leq f(y)$
- **Boundedness**: $(\exists L > 0)(\forall n \in \mathbb{N})\ E(\|g_n\|^2 | x_n) \leq L^2$.

Theorem L23-1: Assuming the previous assumptions hold. Then $E(\mu_k) \to \mu$ as $k \to \infty$, where $\mu_k \coloneqq \min\{f(x_0), \dots, f(x_k)\} \geq \mu$.

Proof: Let $s \in S$ and let $n \in \mathbb{N}$. Then

$$
\begin{aligned}
0 \leq \|x_{n+1} - s\|^2 &= \|P_C(x_n - t_n g_n) - P_C(s)\|^2 \\
&\leq \|(x_n - t_n g_n) - s\|^2 \\
&= \|(x_n - s) - t_n g_n\|^2 \\
&= \|x_n - s\|^2 - 2t_n \langle g_n, x_n - s \rangle + t_n^2 \|g_n\|^2
\end{aligned}
$$

Now taking the conditional expectation, given x_n, yields

$$
\begin{aligned}
E(\|x_{n+1} - s\|^2 | x_n) &\leq \|x_n - s\|^2 + 2t_n \langle E(g_n|x_n), s - x_n \rangle \\
&\quad + t_n^2 E(\|g_n\|^2 | x_n) \\
&\leq \|x_n - s\|^2 + 2t_n (f(s) - f(x_n)) + t_n^2 L^2 \\
&= \|x_n - s\|^2 + 2t_n (\mu - f(x_n)) + t_n^2 L^2
\end{aligned}
$$

Now taking the expectation w.r.t. x_n yields $E(\|x_{n+1} - s\|^2) \leq E(\|x_n - s\|^2) + 2t_n(\mu - E(f(x_n))) + t_n^2 L^2 \quad (\mathrm{I})$. Let $k \in \mathbb{N}$. Summing $\sum_{n=0}^{k}$ over (I) and cancelling duplicate terms yields $0 \leq E(\|x_{n+1} - s\|^2) \leq \|x_0 - s\|^2 - 2\sum_{n=0}^{k} t_n (E(f(x_n)) - \mu) + L^2 \sum_{n=0}^{k} t_n^2$. Hence,

$$
\begin{aligned}
\frac{1}{2}\left(\|x_0 - s\|^2 + L^2 \sum_{n=0}^{k} t_n^2 \right) &\geq \sum_{n=0}^{k} t_n (E(f(x_n)) - \mu) \\
&\geq \sum_{n=0}^{k} t_n (E(\mu_k) - \mu) \\
&\geq 0 \quad \because f(x_n) \geq \mu_k \geq \mu
\end{aligned}
$$

□

Therefore, $0 \leq E(\mu_k) - \mu \leq \dfrac{\|x_0 - s\|^2 + L^2 \sum_{n=0}^{k} t_n^2}{2\sum_{n=0}^{k} t_n} \to 0$ as $k \to +\infty$ by assumption.

Key application: minimizing a sum of functions $f_1, \ldots, f_r : \mathbb{R}^m \to (-\infty, +\infty]$ are convex, l.s.c. and proper. Set $I = \{1, \ldots, r\}$ and assume $(\forall i \in I)\ \mathrm{int}(\mathrm{dom}(f_i)) \supseteq C$ is convex, closed and nonempty. Also assume that $(\forall i \in I)(\exists L_i \geq 0)\ \sup \|\partial f_i(C)\| \leq L_i$.

> **Fact L23-a:** $\sup \|\partial f_i(C)\| \leq L_i \iff f_i|_C$ is L_i-Lipschitz continuous. This is true, for example, C is bounded.

Define $f = \sum_{i \in I} f_i$, and set goal $\min_{x \in C} f(x)$. We will apply SPGM to P. To do that, we can easily verify f is convex, l.s.c. and proper. And

$$
\mathrm{int}(\mathrm{dom}(f)) = \mathrm{int}\left(\bigcap_{i \in I} \mathrm{dom}(f_i) \right) = \bigcap_{i \in I} \mathrm{int}(\mathrm{dom}(f_i)) \supseteq C
$$

Now assume $\mu \coloneqq \min f(C)$ is attained, i.e. P has a solution. we now will show that the assumptions on g_n can both be satisfied. By the Fact L23-a, we have each $f_i|_C$ is L_i-Lipschitz.

Therefore, using the triangle inequality $f|_C = \sum_{i \in I} f_i|_C$ is $\left(\sum_{i \in I} L_i \right)$-Lipschitz. Therefore, once again, by Fact L23-a, we get $\sup \|\partial f(C)\| \leq \sum L_i$.

Let $x_0 \in C$. Given $x_n \in C$, we pick an index $i_n \in I = \{1, \ldots, r\}$ randomly using uniform distribution and we pick a vector $f'_{i_n}(x_n)$ such that $g_n = r \cdot f'_{i_n}(x_n) \in r \cdot \partial f_{i_n}(x_n)$. Now,

$$
\begin{aligned}
E(g_n | x_n) &= \sum_{i=1}^{r} \frac{1}{r} \cdot r f'_i(x_n) \\
&= \sum_{i=1}^{r} f'_i(x_n) \\
&\in \partial f_1(x_n) + \cdots + \partial f_r(x_n) \\
&= \partial(f_1 + \cdots + f_r)(x_n) \quad \text{by sum rule} \\
&= \partial f(x_n)
\end{aligned}
$$

So the unbiased subgradient assumption of g_n holds. Next,

$$\begin{aligned} E(\|g_n\|^2|x_n) &= \sum_{i=1}^{r} \frac{1}{r} \|r f_i'(x_n)\|^2 \\ &= \sum_{i=1}^{r} r\|f_i'(x_n)\|^2 \\ &\leq r\sum L_i^2 =: L^2 \end{aligned}$$

Therefore the boundedness assumption of g_n holds with $L := \sqrt{r\sum L_i^2}$. Consequently, $x_{n+1} := P_C(x_n - t_n g_{i_n})$ generates a sequence such that $E(\mu_n) \to \mu$, where $\mu_n = \min\limits_{i\in\{1,\ldots,n\}} \{f(x_0), \ldots, f(x_n)\}$. □

Duality: The Fenchel Duality

Let $f, g : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Consider the problem $(\mathrm{P}) \quad \min\limits_{x\in\mathbb{R}^m} f(x) + g(x)$. Rewrite the problem as: $\min\limits_{x,z\in\mathbb{R}^m} \{f(x) + g(z) : x = z\}$ and construct the Lagrangian

$$L(x, z, y) := f(x) + g(z) + \langle y, z - x\rangle$$

The dual objective function is obtained by minimizing the Lagrangian w.r.t. x, z.

$$\begin{aligned} d(u) &= \inf_{x,z} L(x, z, u) \\ &= \inf_{x,z} \{f(x) - \langle u, x\rangle + g(z) + \langle u, z\rangle\} \\ &= \inf_{x,z} (-(\langle u, x\rangle - f(x)) - (\langle -u, z\rangle - g(z))) \\ &= -\sup_{x}(\langle u, x\rangle - f(x)) - \sup_{z}(\langle -u, z\rangle - g(z)) \\ &= -f^*(u) - g^*(-u) \end{aligned}$$

We obtain the Fenchel dual problem

$$\begin{aligned} (\mathrm{D}) \quad &\max_{u\in\mathbb{R}^m}(-f^*(u) - g^*(-u)) \\ = &\min_{u\in\mathbb{R}^m}(f^*(u) + g^*(-u)) \end{aligned}$$

Now let

$$\begin{cases} p := \inf\limits_{x\in\mathbb{R}^m} f(x) + g(x) \\ d := \inf\limits_{u\in\mathbb{R}^m} f^*(u) + g^*(-u) \end{cases}$$

By a result from assignment 4, $p \geq -d$.

The Fenchel-Rockafellar Duality

Let $f, g : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Let A be an $n \times m$ matrix, hence $A : \mathbb{R}^m \to \mathbb{R}^n$.

Consider the problem $(\mathrm{P}) \quad \min\limits_{x\in\mathbb{R}^m} f(x) + g(Ax)$ and its Fenchel-Rockafellar dual $(\mathrm{D}) \quad \min\limits_{y\in\mathbb{R}^n} f^*(-A^\intercal y) + g^*(y)$.

As before,

$$\begin{cases} p := \inf\limits_{x\in\mathbb{R}^m} f(x) + g(Ax) \\ d := \inf\limits_{y\in\mathbb{R}^n} f^*(-A^\intercal y) + g^*(y) \end{cases} \implies p \geq -d$$

Lemma L24-a: Let $h : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c., and proper. Set $(\forall x \in \mathbb{R}^m)\ h^v(x) = h(-x)$. Then the following hold:

1. h^v is convex, l.s.c. and proper.
2. $\partial h^v = -\partial h \circ (-\mathrm{Id})$

Proof: for part 1, clearly, $\mathrm{dom}(h^v) = -\mathrm{dom}(h)$, hence $\mathrm{dom}(h^v) \neq \emptyset$. Moreover, $-\infty \notin h(\mathbb{R}^m) = h(-\mathbb{R}^m) = h^v(\mathbb{R}^m) \implies h^v$ is proper.

Now let $x_n \to \bar{x}$ we note $\liminf h^v(x_n) = \liminf h(-x_n) \underbrace{\geq h(-\bar{x})}_{h \text{ is l.s.c.}} = h^v(\bar{x}) \implies h^v$ is l.s.c. Finally, we can show $h^v(\lambda x + (1-\lambda)y) \leq \lambda h^v(x) + (1-\lambda)h^v(y)$, i.e. h^v is convex.

For part 2, let $u \in \mathbb{R}^m$, let $x \in \mathrm{dom}(\partial h \circ (-\mathrm{Id})), u \in -\partial h \circ (-\mathrm{Id})(x) = -\partial h(-x)$

$$\begin{aligned}
&\iff -u \in \partial h(-x) \\
&\iff (\forall y \in \mathbb{R}^m)\ h(y) \geq h(-x) + \langle -u, y + x \rangle \\
&\iff (\forall y \in \mathbb{R}^m)\ h^v(y) \geq h^v(x) + \langle u, y - x \rangle \\
&\iff u \in \partial h^v(x)
\end{aligned}$$

And we're done. □

DR as a Self-Duel Method

Recall that the DR operator to solve P is defined as

$$\begin{aligned}
T_{\text{primal}} &\coloneqq T_p \\
&= \mathrm{Id} - \mathrm{Prox}_f + \mathrm{Prox}_g \circ R_f \\
&= \frac{1}{2}(\mathrm{Id} + R_g \circ R_f)
\end{aligned}$$

where $R_f = 2\mathrm{Prox}_f - \mathrm{Id}$.

Similarly, the DR operator to solve D is defined as

$$\begin{aligned}
T_{\text{dual}} &\coloneqq T_d \\
&= \mathrm{Id} - \mathrm{Prox}_{f^*} + \mathrm{Prox}_{(g^*)^v} \circ R_{f^*} \\
&= \frac{1}{2}
\end{aligned}$$

And $T_p = T_d$! (as we shall show)

Lemma L24-b: Let $h : \mathbb{R}^m \to (-\infty, +\infty]$ be convex, l.s.c. and proper. Set $(\forall x \in \mathbb{R}^m)\ h^v(x) = h(-x)$. Then the following hold:

1. $\mathrm{Prox}_{h^v} = -\mathrm{Prox}_h \circ (-\mathrm{Id})$
2. $R_{h^*} = -R_h$
3. $R_{(h^*)^v} = R_h \circ (-\mathrm{Id})$

Proof: For part 1,

$$\begin{aligned}
\mathrm{Prox}_{h^v} &= (\mathrm{Id} + \partial h^v)^{-1} \\
&= (\mathrm{Id} + (-\mathrm{Id}) \circ \partial h \circ (-\mathrm{Id}))^{-1} \\
&= (-\mathrm{Id})^{-1} \circ (\mathrm{Id} + \partial h)^{-1} \circ (-\mathrm{Id})^{-1} \\
&= -\mathrm{Prox}_h \circ (-\mathrm{Id})
\end{aligned}$$

For part 2,

$$\begin{aligned}
R_{h^*} &\coloneqq 2\mathrm{Prox}_{h^*} - \mathrm{Id} \\
&= 2(\mathrm{Id} - \mathrm{Prox}_h) - \mathrm{Id} \\
&= 2\mathrm{Id} - 2\mathrm{Prox}_h - \mathrm{Id} \\
&= \mathrm{Id} - 2\mathrm{Prox}_h \\
&= -(2\mathrm{Prox}_h - \mathrm{Id}) = -R_h
\end{aligned}$$

For part 3, first note that

$$\begin{aligned}
\mathrm{Prox}_{(h^*)^v} &= -\mathrm{Prox}_{h^*} \circ (-\mathrm{Id}) \\
&= -(\mathrm{Id} - \mathrm{Prox}_h) \circ (-\mathrm{Id}) \\
&= -\mathrm{Id} \circ (-\mathrm{Id}) + \mathrm{Prox}_h \circ (-\mathrm{Id}) \\
&= \mathrm{Prox}_h \circ (-\mathrm{Id}) + \mathrm{Id} \\
&= (\mathrm{Prox}_h - \mathrm{Id}) \circ (-\mathrm{Id})
\end{aligned}$$

Therefore,

$$\begin{aligned}
R_{(h^*)^v} &= 2\mathrm{Prox}_{(h^*)^v} - \mathrm{Id} \\
&= 2(\mathrm{Prox}_h - \mathrm{Id}) \circ (-\mathrm{Id}) - \mathrm{Id} \\
&= (2\mathrm{Prox}_h - 2\mathrm{Id} + \mathrm{Id}) \circ (-\mathrm{Id}) \\
&= R_h \circ (-\mathrm{Id})
\end{aligned}$$

And we're done. □

Theorem L24-c: $T_p = T_d$.

Proof:

$$\begin{aligned} T_d &= \frac{1}{2}(\mathrm{Id} + R_{(g^*)^v} \circ R_{f^*}) \\ &= \frac{1}{2}(\mathrm{Id} + (R_g \circ (-\mathrm{Id})) \circ (-R_f)) \\ &= \frac{1}{2}(\mathrm{Id} + R_g \circ R_f) = T_p \end{aligned}$$

Done! □

Theorem L24-d: Let $x_0 \in \mathbb{R}^m$. Update via

$$x_{n+1} \coloneqq x_n - \mathrm{Prox}_f(x_n) + \mathrm{Prox}_g \circ (2\mathrm{Prox}_f(x_n) - x_n)$$

Then $\mathrm{Prox}_f(x_n) \to$ a minimizer of $f + g$, and $x_n - \mathrm{Prox}_f(x_n) \to$ a minimizer of $f^* + (g^*)^v$.

Proof: Combine Theorem L22-d, with the fact that $T_p = T_d$ to learn that $\mathrm{Prox}_{f^*}(x_n) \to$ a minimizer of $f^* + (g^*)^v$. Now use the fact that $\mathrm{Prox}_{f^*} = \mathrm{Id} - \mathrm{Prox}_f$. □