Optimization for Machine Learning

Gouri Shanker

Contents

1 Vector Spaces

Definition 1.1 (Field)
A *field* is a nonempty set $\mathbb{F}$ together with two binary operations

$$+ : \mathbb{F} \times \mathbb{F} \to \mathbb{F}, \qquad \cdot : \mathbb{F} \times \mathbb{F} \to \mathbb{F},$$

called binary *addition* and binary *multiplication*, satisfying the following axioms.

1. **Additive structure.** The tuple $(\mathbb{F}, +)$ is an abelian group:

   (A1) $(a+b)+c = a+(b+c), \quad \forall a,b,c \in \mathbb{F}$

   (A2) $a+b = b+a, \quad \forall a,b \in \mathbb{F}$

   (A3) $\exists\, 0 \in \mathbb{F}$ such that $a+0 = a, \quad \forall a \in \mathbb{F}$

   (A4) $\forall a \in \mathbb{F},\ \exists\, (-a) \in \mathbb{F}$ such that $a + (-a) = 0$

2. **Multiplicative structure.** The tuple $(\mathbb{F} \setminus \{0\}, \cdot)$ is an abelian group:

   (M1) $(a \cdot b) \cdot c = a \cdot (b \cdot c), \quad \forall a,b,c \in \mathbb{F}$

   (M2) $a \cdot b = b \cdot a, \quad \forall a,b \in \mathbb{F}$

   (M3) $\exists\, 1 \in \mathbb{F}$, $1 \neq 0$, such that $a \cdot 1 = a, \quad \forall a \in \mathbb{F}$

   (M4) $\forall a \in \mathbb{F} \setminus \{0\},\ \exists\, a^{-1} \in \mathbb{F}$ such that $a \cdot a^{-1} = 1$

3. **Compatibility.**

   (D) $a \cdot (b+c) = a \cdot b + a \cdot c, \quad \forall a,b,c \in \mathbb{F}$

Lemma 1.2 (Uniqueness of Inverse)
Let $(\mathbb{F}, +)$ be a group and let $a \in \mathbb{F}$. If $b, c \in \mathbb{F}$ satisfy

$$a+b = 0 \qquad \text{and} \qquad a+c = 0,$$

then $b = c$.

Proof. Using associativity and the additive identity,

$$b = b+0 = b+(a+c).$$

By associativity,

$$b+(a+c) = (b+a)+c.$$

By commutativity, $b+a = a+b = 0$, hence

$$(b+a)+c = 0+c = c.$$

Therefore $b = c$. $\square$

Lemma 1.3 (Uniqueness of Multiplicative Inverse)
Let $(\mathbb{F} \setminus \{0\}, \cdot)$ be a group and let $a \neq 0$. If $b, c \in \mathbb{F}$ satisfy

$$a \cdot b = 1 \qquad \text{and} \qquad a \cdot c = 1,$$

then $b = c$.

Definition 1.4 (Vector Space)
Let $\mathbb{F}$ be a field. A *vector space* over $\mathbb{F}$ is a nonempty set V together with two operations

$$+ : V \times V \to V, \qquad \cdot : \mathbb{F} \times V \to V,$$

called *vector addition* and *scalar multiplication*, such that the following axioms hold.

1. **Additive structure.** The tuple $(V, +)$ is an abelian group:

   (V1) $(\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w})$

   (V2) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$

   (V3) $\exists\, \mathbf{0} \in V$ such that $\mathbf{v} + \mathbf{0} = \mathbf{v}$

   (V4) $\forall \mathbf{v} \in V,\ \exists\, (-\mathbf{v}) \in V$ such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$

2. **Scalar multiplication axioms.** For all $a, b \in \mathbb{F}$ and $\mathbf{u}, \mathbf{v} \in V$:

   (S1) $(ab)\mathbf{v} = a(b\mathbf{v})$

   (S2) $1\mathbf{v} = \mathbf{v}$

   (S3) $a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$

   (S4) $(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$

Example $((\mathbb{F}^n, \mathbb{F}, +, \cdot))$
Let $\mathbb{F}$ be a field and $n \in \mathbb{N}$. Define

$$\mathbb{F}^n := \{(x_1, \ldots, x_n) : x_i \in \mathbb{F}\}.$$

Addition and scalar multiplication are defined componentwise:

$$(x_1, \ldots, x_n) + (y_1, \ldots, y_n) := (x_1 + y_1, \ldots, x_n + y_n),$$

$$a(x_1, \ldots, x_n) := (ax_1, \ldots, ax_n).$$

Then $(\mathbb{F}^n, \mathbb{F}, +, \cdot)$ is a vector space over $\mathbb{F}$.

Example $((\mathbb{F}^{m \times n}, \mathbb{F}, +, \cdot))$

Let $m, n \in \mathbb{N}$. Define

$$\mathbb{F}^{m\times n} := \{A = (a_{ij}) : a_{ij} \in \mathbb{F}\}.$$

Addition and scalar multiplication are defined entrywise:

$$(A+B)_{ij} = a_{ij} + b_{ij}, \qquad (aA)_{ij} = a\, a_{ij}.$$

Then $(\mathbb{F}^{m\times n}, \mathbb{F}, +, \cdot)$ is a vector space over $\mathbb{F}$.

Example $((\mathbb{P}_n, \mathbb{F}, +, \cdot))$
Let $n \in \mathbb{N}$. Define

$$\mathbb{P}_n := \left\{ a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \;\middle|\; a_0, \ldots, a_n \in \mathbb{F} \right\}.$$

Addition and scalar multiplication are defined by

$$(p+q)(x) := p(x) + q(x), \qquad (ap)(x) := a\, p(x).$$

Then $(\mathbb{P}_n, \mathbb{F}, +, \cdot)$ is a vector space over $\mathbb{F}$ of dimension $n+1$.

Definition 1.5 (Linear Combination)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space and let

$$\mathbf{v}_1, \ldots, \mathbf{v}_k \in V, \qquad a_1, \ldots, a_k \in \mathbb{F}.$$

A *linear combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ is a vector of the form

$$a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_k\mathbf{v}_k.$$

Definition 1.6 (Subspace)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space. A subset $W \subseteq V$ is called a *subspace* of V if $(W, \mathbb{F}, +, \cdot)$ is itself a vector space.

Theorem 1.7 (Subspace Criterion)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space and let $S \subseteq V$ be nonempty. Then S is a subspace of V if and only if

$$\alpha u + v \in S \quad \text{for all } \alpha \in \mathbb{F} \text{ and all } u, v \in S.$$

Proof. $(\Rightarrow)$ Assume S is a subspace of V. Then S is closed under vector addition and scalar multiplication. Hence, for any $\alpha \in \mathbb{F}$ and any $u, v \in S$,

$$\alpha u \in S \quad \text{and} \quad \alpha u + v \in S.$$

$(\Leftarrow)$ Conversely, assume $S \subseteq V$ is nonempty and satisfies

$$\alpha u + v \in S \quad \text{for all } \alpha \in \mathbb{F},\ u, v \in S.$$

Closure under scalar multiplication. Fix $u \in S$ and $\alpha \in \mathbb{F}$. Since S is nonempty, choose $v \in S$. Taking $v = 0u = 0 \cdot u$, we obtain

$$\alpha u = \alpha u + 0 \in S.$$

Closure under addition. Let $u, v \in S$. Taking $\alpha = 1$, we have

$$u + v = 1 \cdot u + v \in S.$$

Existence of additive identity. Let $u \in S$. Taking $\alpha = 0$, we obtain

$$0 = 0 \cdot u + u \in S.$$

Existence of additive inverse. Let $u \in S$. Since $0 \in S$, taking $\alpha = -1$ gives

$$-u = (-1)u + 0 \in S.$$

Thus S contains 0, is closed under addition and scalar multiplication, and contains additive inverses. Therefore $(S, +, \cdot)$ is a vector space, and hence S is a subspace of V. □

Theorem 1.8 (Intersection of Subspaces)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space and let $\mathcal{S}$ be a nonempty collection of subspaces of V. Define

$$W := \bigcap_{S \in \mathcal{S}} S.$$

Then W is a subspace of V.

Proof. Since each $S \in \mathcal{S}$ is a subspace, we have $\mathbf{0} \in S$ for all $S \in \mathcal{S}$. Hence $\mathbf{0} \in W$, and thus W is nonempty.

Let $\mathbf{u}, \mathbf{v} \in W$ and let $\alpha \in \mathbb{F}$. Then $\mathbf{u}, \mathbf{v} \in S$ for every $S \in \mathcal{S}$. Since each S is a subspace, it is closed under linear combinations, and therefore

$$\alpha \mathbf{u} + \mathbf{v} \in S \quad \text{for all } S \in \mathcal{S}.$$

Hence $\alpha \mathbf{u} + \mathbf{v} \in \bigcap_{S \in \mathcal{S}} S = W$.

By the subspace criterion, W is a subspace of V. □

Definition 1.9 (Subspace Spanned by a Set)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space and let

$$S := \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subseteq V.$$

Let $\mathcal{S}$ denote the collection of all subspaces of V that contain S, and define

$$W := \bigcap_{K \in \mathcal{S}} K.$$

Then W is a subspace of V, called the *subspace spanned by* S, and is denoted by

$$W = \text{span}(S).$$

Proposition 1.10

$$\text{span}(S) = \left\{ \sum_{i=1}^{n} \alpha_i \mathbf{v}_i \;\middle|\; \alpha_1, \ldots, \alpha_n \in \mathbb{F} \right\}.$$

Proof. Let

$$S = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\} \subseteq V,$$

and let

$$W := \bigcap_{K \in \mathcal{S}} K,$$

where $\mathcal{S}$ denotes the collection of all subspaces of V containing S. Let

$$\text{span}(S) := \left\{ \sum_{i=1}^{n} \alpha_i \mathbf{v}_i \;\middle|\; \alpha_1, \ldots, \alpha_n \in \mathbb{F} \right\}.$$

Step 1: span$(S) \subseteq W$.

Let $\mathbf{u} \in \text{span}(S)$. Then

$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i \quad \text{for some } \alpha_1, \ldots, \alpha_n \in \mathbb{F}.$$

Let $K \in \mathcal{S}$ be arbitrary. Since K is a subspace containing S, we have $\mathbf{v}_i \in K$ for all i. By closure of K under linear combinations,

$$\mathbf{u} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i \in K.$$

Since this holds for every $K \in \mathcal{S}$, it follows that

$$\mathbf{u} \in \bigcap_{K \in \mathcal{S}} K = W.$$

Hence span$(S) \subseteq W$.

Step 2: $W \subseteq$ span(S).

We first note that span(S) is a subspace of V and contains S. Therefore,

$$\text{span}(S) \in \mathcal{S}.$$

By definition of W as the intersection of all elements of $\mathcal{S}$,

$$W = \bigcap_{K \in \mathcal{S}} K \subseteq \text{span}(S).$$

Combining the two inclusions, we conclude that

$$W = \text{span}(S).$$

□

Definition 1.11 (Linear Independence)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space and let

$$S := \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subseteq V.$$

The set S is said to be *linearly independent* if

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_n \mathbf{v}_n = \mathbf{0}$$

implies

$$\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0.$$

Otherwise, S is called *linearly dependent*.

Definition 1.12 (Basis)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space and let

$$S := \{\mathbf{v}_1, \dots, \mathbf{v}_n\} \subseteq V.$$

The set S is called a *basis* of V if

1. S is linearly independent, and
2. $\text{span}(S) = V$.

Remark
A vector space $(V, \mathbb{F}, +, \cdot)$ is said to be *finite-dimensional* if there exists a finite set $B \subseteq V$ that forms a basis of V.

Remark
If $S = \{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is a basis of V, then the integer n is called the *dimension* of V and is denoted by $\dim V = n$.

Theorem 1.13
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space and suppose

$$V = \text{span}(S), \qquad S = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}.$$

Then any linearly independent set of vectors in V is finite and contains at most m vectors.

Proof. Let

$$L = \{\mathbf{u}_1, \dots, \mathbf{u}_k\} \subseteq V$$

be a linearly independent set. Since $V = \text{span}(S)$, each $\mathbf{u}_j$ can be written as a linear combination of $\mathbf{v}_1, \dots, \mathbf{v}_m$:

$$\mathbf{u}_j = \sum_{i=1}^{m} a_{ij} \mathbf{v}_i, \qquad a_{ij} \in \mathbb{F}.$$

Suppose, for contradiction, that $k > m$. Consider a linear combination

$$\alpha_1 \mathbf{u}_1 + \cdots + \alpha_k \mathbf{u}_k = \mathbf{0}.$$

Substituting the expressions for $\mathbf{u}_j$,

$$\sum_{j=1}^{k} \alpha_j \left(\sum_{i=1}^{m} a_{ij} \mathbf{v}_i \right) = \sum_{i=1}^{m} \left(\sum_{j=1}^{k} \alpha_j a_{ij} \right) \mathbf{v}_i = \mathbf{0}.$$

This is a linear combination of the m vectors $\mathbf{v}_1, \dots, \mathbf{v}_m$. Since there are $k > m$ scalars $\alpha_1, \dots, \alpha_k$, the homogeneous system

$$\sum_{j=1}^{k} \alpha_j a_{ij} = 0, \qquad i = 1, \dots, m,$$

has a nontrivial solution. Hence there exist scalars, not all zero, such that

$$\alpha_1 \mathbf{u}_1 + \cdots + \alpha_k \mathbf{u}_k = \mathbf{0},$$

contradicting the linear independence of L.

Therefore $k \leq m$. Thus every linearly independent set in V is finite and contains no more than m vectors. □

Corollary 1.14
Let $(V, \mathbb{F}, +, \cdot)$ be a finite-dimensional vector space. Then any two bases of V contain the same number of vectors.

Proof. Let

$$B_1 = \{\mathbf{v}_1, \dots, \mathbf{v}_m\}, \qquad B_2 = \{\mathbf{u}_1, \dots, \mathbf{u}_n\}$$

be two bases of V. Since B_1 spans V and B_2 is linearly independent, by Theorem 1.13 we have

$$n \leq m.$$

Similarly, since B_2 spans V and B_1 is linearly independent, the same theorem gives

$$m \leq n.$$

Therefore $m = n$. □

Definition 1.15 (Linear Transformation)
Let $(V, \mathbb{F}, +, \cdot)$ and $(W, \mathbb{F}, +, \cdot)$ be vector spaces. A map

$$T : V \to W$$

is called a *linear transformation* if for all $\mathbf{u}, \mathbf{v} \in V$ and all $\alpha \in \mathbb{F}$,
(L1) $T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v})$ (additivity)
(L2) $T(\alpha \mathbf{u}) = \alpha\, T(\mathbf{u})$ (homogeneity)

Remark
A map $T : V \to W$ is linear if and only if

$$T(\alpha \mathbf{u} + \mathbf{v}) = \alpha T(\mathbf{u}) + T(\mathbf{v}) \quad \forall \alpha \in \mathbb{F},\ \mathbf{u}, \mathbf{v} \in V.$$

Proposition 1.16
If $T : V \to W$ is linear, then

$$T(\mathbf{0}) = \mathbf{0}, \qquad T(-\mathbf{u}) = -T(\mathbf{u}) \quad \forall\, \mathbf{u} \in V.$$

Definition 1.17 (Vector Space Homomorphism)
Let $(V, \mathbb{F}, +, \cdot)$ and $(W, \mathbb{F}, +, \cdot)$ be vector spaces. A map

$$T : V \to W$$

is called a *(vector space) homomorphism* if for all $\mathbf{u}, \mathbf{v} \in V$ and all $\alpha \in \mathbb{F}$,

$$T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v}), \qquad T(\alpha \mathbf{u}) = \alpha\, T(\mathbf{u}).$$

Remark
A map $T : V \to W$ is a homomorphism if and only if

$$T(\alpha \mathbf{u} + \mathbf{v}) = \alpha T(\mathbf{u}) + T(\mathbf{v}) \quad \forall\, \alpha \in \mathbb{F},\ \mathbf{u}, \mathbf{v} \in V.$$

Definition 1.18 (Linear Functional)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space. A *linear functional* on V is a linear transformation

$$f : V \to \mathbb{F}.$$

That is, for all $\mathbf{u}, \mathbf{v} \in V$ and all $\alpha \in \mathbb{F}$,

$$f(\mathbf{u} + \mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v}), \qquad f(\alpha \mathbf{u}) = \alpha\, f(\mathbf{u}).$$

Example (Evaluation Functional)
Fix $a \in \mathbb{F}$. Define $f_a : \mathbb{F}[x] \to \mathbb{F}$ by

$$f_a(p) := p(a).$$

Then f_a is a linear functional.

Example (Constant-Coefficient Functional)
Define $f_0 : \mathbb{P}_n \to \mathbb{F}$ by

$$f_0(a_0 + a_1 x + \cdots + a_n x^n) := a_0.$$

Then f_0 is a linear functional.

Remark
The set of all linear functionals on V forms a vector space over $\mathbb{F}$, called the *dual space* and denoted by V^*.

Definition 1.19 (Inner Product)
Let V be a vector space over $\mathbb{F}$, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. An *inner product*

on V is a map

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$$

satisfying the following properties for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$ and all $\alpha \in \mathbb{F}$:

(IP1) $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$ (conjugate symmetry)

(IP2) $\langle \alpha \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \alpha \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$ (linearity in the first argument)

(IP3) $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ and $\langle \mathbf{u}, \mathbf{u} \rangle = 0 \iff \mathbf{u} = \mathbf{0}$ (positive definiteness)

If $\mathbb{F} = \mathbb{R}$, conjugation is trivial and symmetry reduces to $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$

Definition 1.20 (Inner Product Space)
An *inner product space* is a pair $(V, \langle \cdot, \cdot \rangle)$ where V is a vector space over $\mathbb{F}$ and $\langle \cdot, \cdot \rangle$ is an inner product on V.

Example ($\mathbb{F}^n$)

$$\langle \mathbf{x}, \mathbf{y} \rangle := \sum_{i=1}^{n} x_i \overline{y_i}.$$

Example ($\mathbb{P}_n$)

$$\langle p, q \rangle := \int_0^1 p(x) \overline{q(x)} \, dx.$$

Proposition 1.21
For all $\mathbf{u}, \mathbf{v} \in V$:

$$\langle \mathbf{u}, \mathbf{v} \rangle = 0 \iff \mathbf{u} \perp \mathbf{v}, \qquad \|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}.$$

Definition 1.22 (Norm)
For $\mathbf{u} \in V$, define

$$\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}.$$

Definition 1.23 (Orthogonality)
Vectors $\mathbf{u}, \mathbf{v} \in V$ are said to be *orthogonal*, written $\mathbf{u} \perp \mathbf{v}$, if

$$\langle \mathbf{u}, \mathbf{v} \rangle = 0.$$

Proposition 1.24
Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. Then for all $\mathbf{u}, \mathbf{v} \in V$:

1. $\|\mathbf{u}\| \geq 0$ and $\|\mathbf{u}\| = 0 \iff \mathbf{u} = \mathbf{0}$
2. $\|\alpha \mathbf{u}\| = |\alpha| \, \|\mathbf{u}\|$
3. $\langle \mathbf{u}, \mathbf{v} \rangle = 0 \iff \mathbf{u} \perp \mathbf{v}$

Definition 1.25 (Norm)
Let $(V, \mathbb{F}, +, \cdot)$ be a vector space, where $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$. A *norm* on V is a function

$$\|\cdot\| : V \to [0, \infty)$$

satisfying, for all $\mathbf{u}, \mathbf{v} \in V$ and all $\alpha \in \mathbb{F}$,
(N1) $\|\mathbf{u}\| \geq 0$ and $\|\mathbf{u}\| = 0$ if and only if $\mathbf{u} = \mathbf{0}$ (positive definiteness)
(N2) $\|\alpha\mathbf{u}\| = |\alpha| \, \|\mathbf{u}\|$ (absolute homogeneity)
(N3) $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$ (triangle inequality)

Definition 1.26 (Normed Vector Space)
A *normed vector space* is a pair $(V, \|\cdot\|)$, where V is a vector space and $\|\cdot\|$ is a norm on V.

Example (ℓ^p-type norms on $\mathbb{F}^n$)
For $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{F}^n$, define

$$\|\mathbf{x}\|_p := \begin{cases} \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}, & 1 \leq p < \infty, \\ \max_{1 \leq i \leq n} |x_i|, & p = \infty. \end{cases}$$

Then $(\mathbb{F}^n, \|\cdot\|_p)$ is a normed vector space.

Example (Norm induced by an inner product)
Let $(V, \langle \cdot, \cdot \rangle)$ be an inner product space. Define

$$\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}.$$

Then $\|\cdot\|$ is a norm on V.

Example (Polynomial norm)
Let $V = \mathbb{P}_n$ and $\mathbb{F} = \mathbb{R}$. Define

$$\|p\| := \max_{x \in [0,1]} |p(x)|.$$

Then $(\mathbb{P}_n, \|\cdot\|)$ is a normed vector space.

Proposition 1.27
Let $(V, \|\cdot\|)$ be a normed vector space. Then for all $\mathbf{u}, \mathbf{v} \in V$:
1. $\|\mathbf{u} - \mathbf{v}\| = 0 \iff \mathbf{u} = \mathbf{v}$
2. $\|\mathbf{u} - \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$
3. $|\|\mathbf{u}\| - \|\mathbf{v}\|| \leq \|\mathbf{u} - \mathbf{v}\|$

Remark
Every inner product induces a norm, but not every norm arises from an inner product.

Definition 1.28 (Metric)
Let X be a nonempty set. A *metric* on X is a function

$$d : X \times X \to [0, \infty)$$

satisfying, for all $x, y, z \in X$:

(M1) $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$ (non-negativity and identity)

(M2) $d(x, y) = d(y, x)$ (symmetry)

(M3) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

Definition 1.29 (Metric Space)
A *metric space* is a pair (X, d), where X is a set and d is a metric on X.

Example (Euclidean Metric)
Let $X = \mathbb{F}^n$ and define

$$d(\mathbf{x}, \mathbf{y}) := \|\mathbf{x} - \mathbf{y}\|_2.$$

Then $(\mathbb{F}^n, d)$ is a metric space.

Example (Discrete Metric)
Let X be any set and define

$$d(x, y) := \begin{cases} 0, & x = y, \\ 1, & x \neq y. \end{cases}$$

Then (X, d) is a metric space.

Example (Polynomial Metric)
Let $X = \mathbb{P}_n$ and define

$$d(p, q) := \|p - q\|_\infty = \max_{x \in [0,1]} |p(x) - q(x)|.$$

Then $(\mathbb{P}_n, d)$ is a metric space.

Proposition 1.30
Let $(V, \|\cdot\|)$ be a normed vector space. Define

$$d(\mathbf{u}, \mathbf{v}) := \|\mathbf{u} - \mathbf{v}\|.$$

Then d is a metric on V.

Proof. Non-negativity and symmetry follow directly from properties of the norm. Moreover,

$$d(\mathbf{u}, \mathbf{v}) = 0 \iff \|\mathbf{u} - \mathbf{v}\| = 0 \iff \mathbf{u} = \mathbf{v}.$$

Finally, by the triangle inequality for the norm,

$$d(\mathbf{u}, \mathbf{w}) = \|\mathbf{u} - \mathbf{w}\| \leq \|\mathbf{u} - \mathbf{v}\| + \|\mathbf{v} - \mathbf{w}\| = d(\mathbf{u}, \mathbf{v}) + d(\mathbf{v}, \mathbf{w}).$$

□

Remark
Every normed vector space is a metric space, but a metric space need not carry any linear or vector space structure.

2 Definitions

The following definitions and properties establish the topological groundwork on $\mathbb{R}^n$ required for convex analysis.

Definition 2.1 (Open Ball)
Fix a norm $\|\cdot\|$ on $\mathbb{R}^n$. An **open ball** of radius $\epsilon > 0$ centered at $\mathbf{x}$, denoted $B_\epsilon(\mathbf{x})$, is defined as:

$$B_\epsilon(\mathbf{x}) := \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{y}\| < \epsilon\}$$

Definition 2.2 (Open Set)
A set $C \subseteq \mathbb{R}^n$ is said to be **open** if for every $\mathbf{x} \in C$, there exists some $\epsilon > 0$ such that the ball is fully contained in the set:

$$B_\epsilon(\mathbf{x}) \subseteq C$$

Definition 2.3 (Closed Set)
A set $C \subseteq \mathbb{R}^n$ is **closed** if its complement $C^c = \mathbb{R}^n \setminus C$ is an open set.

Sequences and Convergence

Definition 2.4 (Sequence)
A sequence is a function from the natural numbers $\mathbb{N}$ to a set X (here $\mathbb{R}^n$). It is denoted as $(\mathbf{x}_k)_{k\geq 1} \subseteq \mathbb{R}^n$.

Definition 2.5 (Convergence)
A sequence $(\mathbf{x}_k)_{k\geq 1}$ is said to **converge** to a limit $\mathbf{x} \in \mathbb{R}^n$ (denoted $\mathbf{x}_k \to \mathbf{x}$) if for every $\epsilon > 0$, there exists an integer $N \in \mathbb{N}$ such that:

$$\|\mathbf{x}_k - \mathbf{x}\| < \epsilon \quad \text{for all } k \geq N$$

Equivalently, $\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}\| = 0$.

Proposition 2.6 (Sequential Characterization of Closed Sets)
A subset $C \subseteq \mathbb{R}^n$ is closed if and only if for every convergent sequence $(\mathbf{x}_k)_{k\geq 1} \subseteq C$ such that $\mathbf{x}_k \to \mathbf{x}$, the limit point satisfies $\mathbf{x} \in C$.

Proof. ($\Rightarrow$) Suppose C is closed and let $(\mathbf{x}_k) \subseteq C$ converge to $\mathbf{x}$. We must show $\mathbf{x} \in C$. Assume for the sake of contradiction that $\mathbf{x} \notin C$. Then $\mathbf{x} \in C^c$. Since C is closed, C^c is open. Therefore, there exists $\epsilon > 0$ such that $B_\epsilon(\mathbf{x}) \subseteq C^c$. Since $\mathbf{x}_k \to \mathbf{x}$, there exists N such that for all $k \geq N$, $\|\mathbf{x}_k - \mathbf{x}\| < \epsilon$, which implies $\mathbf{x}_k \in B_\epsilon(\mathbf{x})$. This means $\mathbf{x}_k \in C^c$, which contradicts the fact that the sequence is contained in C. Thus, $\mathbf{x}$ must belong to C.

($\Leftarrow$) We show that if the limit property holds, C is closed (i.e., C^c is open). Suppose C^c is not open. Then there exists $\mathbf{x} \in C^c$ such that for every $\epsilon > 0$, $B_\epsilon(\mathbf{x}) \not\subseteq C^c$. This implies $B_\epsilon(\mathbf{x}) \cap C \neq \emptyset$. For each $k \geq 1$, choose $\mathbf{x}_k \in B_{1/k}(\mathbf{x}) \cap C$. Then $(\mathbf{x}_k) \subseteq C$ and $\|\mathbf{x}_k - \mathbf{x}\| < 1/k \to 0$, so $\mathbf{x}_k \to \mathbf{x}$. By the hypothesis, the limit $\mathbf{x}$ must be in C. But we started with $\mathbf{x} \in C^c$, a contradiction. □

Interior, Closure, and Boundary

Definition 2.7 (Interior)
The **interior** of a set C, denoted $\text{int}(C)$, is the set of all interior points:

$$\text{int}(C) = \{\mathbf{x} \in C : \exists \epsilon > 0 \text{ such that } B_\epsilon(\mathbf{x}) \subseteq C\}$$

Proposition 2.8
For any set $C \subseteq \mathbb{R}^n$, the interior $\text{int}(C)$ is an open set.

Proof. Let $\mathbf{x} \in \text{int}(C)$. By definition, there exists $\epsilon > 0$ such that $B_\epsilon(\mathbf{x}) \subseteq C$. We claim that $B_\epsilon(\mathbf{x}) \subseteq \text{int}(C)$. Let $\mathbf{y} \in B_\epsilon(\mathbf{x})$. Then $\|\mathbf{x} - \mathbf{y}\| < \epsilon$. Let $\delta = \epsilon - \|\mathbf{x} - \mathbf{y}\| > 0$. We show that $B_\delta(\mathbf{y}) \subseteq C$. Let $\mathbf{z} \in B_\delta(\mathbf{y})$. Then:

$$\|\mathbf{z} - \mathbf{x}\| \leq \|\mathbf{z} - \mathbf{y}\| + \|\mathbf{y} - \mathbf{x}\| < \delta + \|\mathbf{y} - \mathbf{x}\| = \epsilon$$

Thus $\mathbf{z} \in B_\epsilon(\mathbf{x}) \subseteq C$. This shows $B_\delta(\mathbf{y}) \subseteq C$, so $\mathbf{y} \in \text{int}(C)$. Since $\mathbf{y}$ was arbitrary, $B_\epsilon(\mathbf{x}) \subseteq \text{int}(C)$, proving $\text{int}(C)$ is open. □

Definition 2.9 (Closure)
The **closure** of a set C, denoted $\overline{C}$, is the set of all points that are "close" to C:

$$\overline{C} = \{\mathbf{x} \in \mathbb{R}^n : B_\epsilon(\mathbf{x}) \cap C \neq \emptyset, \forall \epsilon > 0\}$$

Intuitively, this set contains all elements of C plus its boundary.

Definition 2.10 (Boundary)
The **boundary** of a set C, denoted ∂C, is the set difference between the closure and the interior:

$$\partial C := \overline{C} \setminus \text{int}(C)$$

Definition 2.11 (Continuity at a point)
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and let $\mathbf{x}_0 \in \operatorname{dom} f$. The function f is said to be *continuous at* $\mathbf{x}_0$ if

$$\lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x}) = f(\mathbf{x}_0),$$

that is, for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\|\mathbf{x} - \mathbf{x}_0\| < \delta \implies |f(\mathbf{x}) - f(\mathbf{x}_0)| < \varepsilon.$$

Convex Sets and Functions

Definition 2.12 (Convex Set)
A set $C \subseteq \mathbb{R}^n$ is called *convex* if for all $\mathbf{x}, \mathbf{y} \in C$ and all $\lambda \in [0,1]$,

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in C.$$

Definition 2.13 (Convex Function)
Let $C \subseteq \mathbb{R}^n$ be a convex set. A function $f : C \to \mathbb{R}$ is called *convex* if for all $\mathbf{x}, \mathbf{y} \in C$ and all $\lambda \in [0,1]$,

$$f\big(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\big) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

Definition 2.14 (Convex Program)
A *convex program* is an optimization problem of the form

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

where $C \subseteq \mathbb{R}^n$ is a convex set and $f : C \to \mathbb{R}$ is a convex function.

Definition 2.15 (Linear Program)
A *linear program* is an optimization problem of the form

$$\min_{\mathbf{x} \in \mathbb{R}^n} \ \mathbf{c}^\mathsf{T} \mathbf{x} \quad \text{subject to } A\mathbf{x} \leq \mathbf{b},$$

where $\mathbf{c} \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, and $\mathbf{b} \in \mathbb{R}^m$.

Remark
Every linear program is a convex program.

Definition 2.16 (Affine Set)
A set $A \subseteq \mathbb{R}^n$ is called *affine* if for all $\mathbf{x}, \mathbf{y} \in A$ and all $\lambda \in \mathbb{R}$,

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{y} \in A.$$

Definition 2.17 (Affine Hull)
Let $S \subseteq \mathbb{R}^n$. The *affine hull* of S, denoted $\operatorname{aff}(S)$, is the smallest affine set containing S, equivalently

$$\operatorname{aff}(S) = \left\{ \sum_{i=1}^{k} \alpha_i \mathbf{x}_i \;\middle|\; \mathbf{x}_i \in S,\ \sum_{i=1}^{k} \alpha_i = 1 \right\}.$$

Definition 2.18 (Graph)
Let $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$. The *graph* of f is the set

$$\operatorname{graph}(f) := \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} : t = f(\mathbf{x}),\ \mathbf{x} \in C\}.$$

Definition 2.19 (Epigraph)
Let $f : C \subseteq \mathbb{R}^n \to \mathbb{R}$. The *epigraph* of f is the set

$$\text{epi}(f) := \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} : t \geq f(\mathbf{x}),\ \mathbf{x} \in C\}.$$

Definition 2.20 (Subgradient)
A vector $\mathbf{g} \in \mathbb{R}^n$ is a **subgradient** of a convex function f at $\mathbf{x}_0$ if:

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{g}^T(\mathbf{x} - \mathbf{x}_0) \quad \forall \mathbf{x} \in \mathsf{dom}(f)$$

The set of all subgradients at $\mathbf{x}_0$ is denoted $\partial f(\mathbf{x}_0)$.

3 Convex Optimization

Theorem 3.1
Let $C \subseteq \mathbb{R}^n$ be a convex set and let $f : C \to \mathbb{R}$ be a function. The function f is convex if and only if its epigraph $\text{epi}(f)$ is a convex set.

Proof. The proof consists of two directions.

Direction 1: $\text{epi}(f)$ is convex $\implies$ f is convex
Assume that $\text{epi}(f)$ is convex. Let $\mathbf{x}, \mathbf{y} \in C$ and let $\lambda \in [0, 1]$. By the definition of the epigraph:

$$(\mathbf{x}, f(\mathbf{x})) \in \text{epi}(f) \quad \text{and} \quad (\mathbf{y}, f(\mathbf{y})) \in \text{epi}(f)$$

Since $\text{epi}(f)$ is a convex set, the convex combination of these two points must also belong to $\text{epi}(f)$:

$$\lambda(\mathbf{x}, f(\mathbf{x})) + (1 - \lambda)(\mathbf{y}, f(\mathbf{y})) \in \text{epi}(f)$$

Expanding this, we get:

$$(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}, \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y})) \in \text{epi}(f)$$

By the definition of the epigraph, if a point $(\mathbf{z}, t) \in \text{epi}(f)$, then $f(\mathbf{z}) \leq t$. Applying this here:

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

Since $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in [0, 1]$ were arbitrary, f is a convex function on C.

Direction 2: f is convex $\implies$ $\text{epi}(f)$ is convex
Assume f is convex. Let $(\mathbf{x}, t_1) \in \text{epi}(f)$ and $(\mathbf{y}, t_2) \in \text{epi}(f)$. This implies:

$$\mathbf{x}, \mathbf{y} \in C, \quad f(\mathbf{x}) \leq t_1, \quad \text{and} \quad f(\mathbf{y}) \leq t_2$$

Consider a convex combination of these points for any $\lambda \in [0, 1]$:

$$\mathbf{P} = \lambda(\mathbf{x}, t_1) + (1 - \lambda)(\mathbf{y}, t_2) = (\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}, \lambda t_1 + (1 - \lambda)t_2)$$

Since C is a convex set, $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in C$. Now we check the convexity condition for f:

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

Because $f(\mathbf{x}) \leq t_1$ and $f(\mathbf{y}) \leq t_2$, it follows that:

$$\lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}) \leq \lambda t_1 + (1 - \lambda)t_2$$

By transitivity:

$$f(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda t_1 + (1 - \lambda)t_2$$

This confirms that the point $\mathbf{P}$ satisfies the definition of the epigraph. Thus, $(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}, \lambda t_1 + (1 - \lambda)t_2) \in \text{epi}(f)$, proving that $\text{epi}(f)$ is a convex set. $\square$

Theorem 3.2 (Strict Separation for Closed/Compact Sets)
Let C and D be two convex subsets of $\mathbb{R}^n$ that are both closed. Furthermore, assume that at least one of them is bounded (compact). If $C \cap D = \emptyset$, then there exists a hyperplane that strictly separates C and D.

Proof. Since C and D are disjoint, closed, and at least one is bounded, the distance between the sets is strictly positive. We define the distance as:

$$\text{dist}(C, D) := \inf\{\|\mathbf{x} - \mathbf{y}\|_2 : \mathbf{x} \in C, \mathbf{y} \in D\}$$

Because of the closed/bounded assumptions, this infimum is attained. There exist points $\mathbf{c} \in C$ and $\mathbf{d} \in D$ such that:

$$\|\mathbf{c} - \mathbf{d}\| = \text{dist}(C, D) > 0$$

Construction of the Hyperplane

We define the normal vector $\mathbf{a}$ and the bias scalar b as follows (using the midpoint construction):

$$\begin{aligned} \mathbf{a} &= \mathbf{d} - \mathbf{c} \\ b &= \frac{\|\mathbf{d}\|^2 - \|\mathbf{c}\|^2}{2} = \mathbf{a}^T \left(\frac{\mathbf{d} + \mathbf{c}}{2} \right) \end{aligned}$$

The separating hyperplane is $H = \{\mathbf{x} : \mathbf{a}^T\mathbf{x} = b\}$. We claim that for all $\mathbf{x} \in C$, $\mathbf{a}^T\mathbf{x} \le \mathbf{a}^T\mathbf{c}$, and for all $\mathbf{y} \in D$, $\mathbf{a}^T\mathbf{y} \ge \mathbf{a}^T\mathbf{d}$.

Proof by Contradiction (Variational Argument)

We show the first part: $\mathbf{a}^T\mathbf{x} \le \mathbf{a}^T\mathbf{c}$ for all $\mathbf{x} \in C$.

Suppose for the sake of contradiction that there exists some $\mathbf{x} \in C$ such that $\mathbf{a}^T\mathbf{x} > \mathbf{a}^T\mathbf{c}$. Consider the line segment connecting $\mathbf{c}$ and $\mathbf{x}$. Since C is convex, the point $\mathbf{z}(t)$ belongs to C for any $t \in [0, 1]$:

$$\mathbf{z}(t) = (1 - t)\mathbf{c} + t\mathbf{x} = \mathbf{c} + t(\mathbf{x} - \mathbf{c})$$

We analyze the squared distance from $\mathbf{d}$ to this point $\mathbf{z}(t)$:

$$\begin{aligned} \|\mathbf{z}(t) - \mathbf{d}\|^2 &= \|\mathbf{c} + t(\mathbf{x} - \mathbf{c}) - \mathbf{d}\|^2 \\ &= \|(\mathbf{c} - \mathbf{d}) + t(\mathbf{x} - \mathbf{c})\|^2 \\ &= \|\mathbf{c} - \mathbf{d}\|^2 + 2t(\mathbf{c} - \mathbf{d})^T(\mathbf{x} - \mathbf{c}) + t^2\|\mathbf{x} - \mathbf{c}\|^2 \end{aligned}$$

Recall that $\mathbf{a} = \mathbf{d} - \mathbf{c}$, so $\mathbf{c} - \mathbf{d} = -\mathbf{a}$. Substituting this back:

$$\|\mathbf{z}(t) - \mathbf{d}\|^2 = \|\mathbf{c} - \mathbf{d}\|^2 - 2t\mathbf{a}^T(\mathbf{x} - \mathbf{c}) + t^2\|\mathbf{x} - \mathbf{c}\|^2$$

From our contradiction assumption, $\mathbf{a}^T\mathbf{x} > \mathbf{a}^T\mathbf{c}$, which implies $\mathbf{a}^T(\mathbf{x} - \mathbf{c}) > 0$. Let $K = \mathbf{a}^T(\mathbf{x} - \mathbf{c}) > 0$.

$$\|\mathbf{z}(t) - \mathbf{d}\|^2 = \|\mathbf{c} - \mathbf{d}\|^2 - t\left[2K - t\|\mathbf{x} - \mathbf{c}\|^2\right]$$

Since $K > 0$, we can choose t strictly positive but small enough (specifically $0 < t < \frac{2K}{\|\mathbf{x}-\mathbf{c}\|^2}$) such that the term in the brackets is positive.

This implies:

$$\|\mathbf{z}(t) - \mathbf{d}\|^2 < \|\mathbf{c} - \mathbf{d}\|^2$$

This contradicts the definition that $\mathbf{c}$ is the closest point in C to D. Therefore, such an $\mathbf{x}$ cannot exist, and we must have $\mathbf{a}^T\mathbf{x} \leq \mathbf{a}^T\mathbf{c}$ for all $\mathbf{x} \in C$.

Final Separation

A symmetric argument shows that $\mathbf{a}^T\mathbf{y} \geq \mathbf{a}^T\mathbf{d}$ for all $\mathbf{y} \in D$. Combining these results with the definition of b:

$$b - \mathbf{a}^T\mathbf{c} = \frac{\|\mathbf{d}-\mathbf{c}\|^2}{2} > 0 \quad \implies \quad \mathbf{a}^T\mathbf{c} < b$$

$$\mathbf{a}^T\mathbf{d} - b = \frac{\|\mathbf{d}-\mathbf{c}\|^2}{2} > 0 \quad \implies \quad b < \mathbf{a}^T\mathbf{d}$$

Thus, for all $\mathbf{x} \in C$ and $\mathbf{y} \in D$, we have the strict separation:

$$\mathbf{a}^T\mathbf{x} \leq \mathbf{a}^T\mathbf{c} < b < \mathbf{a}^T\mathbf{d} \leq \mathbf{a}^T\mathbf{y}$$

□

Theorem 3.3
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then f is continuous on $\mathbb{R}^n$.

Proof. Let $\mathbf{x}_0 \in \mathbb{R}^n$ be arbitrary but fixed. We proceed by showing f is continuous at $\mathbf{x}_0$ using a sequence of claims.

Claim 3.3.1
f is locally bounded in the neighborhood $B_r(\mathbf{x}_0)$.

Let $\mathbf{v}_0, \ldots, \mathbf{v}_n$ be an affinely independent set of points in $\mathbb{R}^n$ such that $\mathbf{x}_0$ lies in the interior of their convex hull, $\text{conv}(\{\mathbf{v}_0, \ldots, \mathbf{v}_n\})$. There exists $r > 0$ such that $B_r(\mathbf{x}_0) \subset \text{conv}(\{\mathbf{v}_0, \ldots, \mathbf{v}_n\})$. Any $\mathbf{x} \in B_r(\mathbf{x}_0)$ can be written as $\mathbf{x} = \sum \alpha_i \mathbf{v}_i$ with $\sum \alpha_i = 1, \alpha_i \geq 0$. Since f is convex:

$$f(\mathbf{x}) \leq \sum \alpha_i f(\mathbf{v}_i) \leq \max_i f(\mathbf{v}_i) := M$$

To bound from below, for any $\mathbf{x} \in B_r(\mathbf{x}_0)$, let $\mathbf{z} = 2\mathbf{x}_0 - \mathbf{x}$. Then $\mathbf{x}_0 = \frac{1}{2}\mathbf{z} + \frac{1}{2}\mathbf{x}$. Since $\mathbf{z} \in B_r(\mathbf{x}_0)$, we have:

$$f(\mathbf{x}_0) \leq \frac{1}{2}f(\mathbf{z}) + \frac{1}{2}f(\mathbf{x}) \implies f(\mathbf{x}) \geq 2f(\mathbf{x}_0) - f(\mathbf{z}) \geq 2f(\mathbf{x}_0) - M := m$$

Thus, $m \leq f(\mathbf{x}) \leq M$ for all $\mathbf{x} \in B_r(\mathbf{x}_0)$.

Claim 3.3.2 (Lipschitz Bound)
If $\phi : [a,b] \to \mathbb{R}$ is convex and bounded ($m \leq \phi(t) \leq M$), then for $s,t \in [a+\delta, b-\delta]$:

$$\frac{|\phi(t)-\phi(s)|}{|t-s|} \leq \frac{M-m}{\delta}$$

Proof. Let $a < a+\delta \leq s < t \leq b-\delta < b$. The property of convex functions states that the slope of the chord $S(x,y) = \dfrac{\phi(y)-\phi(x)}{y-x}$ is non-decreasing in both variables.

We compare the slope on $[s,t]$ with the slopes involving boundary points:

1. Comparing with the left boundary a: Since $a < s < t$, we have $S(a,s) \leq S(s,t)$.

   $$\tfrac{\phi(s)-\phi(a)}{s-a} \leq \tfrac{\phi(t)-\phi(s)}{t-s}$$

   Since $s \geq a+\delta$, then $s-a \geq \delta$. Also $\phi(s) \geq m$ and $\phi(a) \leq M$, so $\phi(s)-\phi(a) \geq m-M$.

   $$\tfrac{m-M}{\delta} \leq \tfrac{\phi(s)-\phi(a)}{s-a} \leq \tfrac{\phi(t)-\phi(s)}{t-s}$$

2. Comparing with the right boundary b: Since $s < t < b$, we have $S(s,t) \leq S(t,b)$.

   $$\tfrac{\phi(t)-\phi(s)}{t-s} \leq \tfrac{\phi(b)-\phi(t)}{b-t}$$

   Since $t \leq b-\delta$, then $b-t \geq \delta$. Also $\phi(b) \leq M$ and $\phi(t) \geq m$, so $\phi(b)-\phi(t) \leq M-m$.

   $$\tfrac{\phi(t)-\phi(s)}{t-s} \leq \tfrac{\phi(b)-\phi(t)}{b-t} \leq \tfrac{M-m}{\delta}$$

Combining both inequalities:

$$-\tfrac{M-m}{\delta} \leq \tfrac{\phi(t)-\phi(s)}{t-s} \leq \tfrac{M-m}{\delta}$$

Taking the absolute value gives the desired linear growth bound:

$$\tfrac{|\phi(t)-\phi(s)|}{|t-s|} \leq \tfrac{M-m}{\delta}$$

□

Claim 3.3.3
If $\varphi : [a,b] \to \mathbb{R}$ is convex and bounded ($m \leq \varphi(t) \leq M$), then for

$s, t \in [a+\delta, b-\delta]$:

$$\frac{|\varphi(t) - \varphi(s)|}{|t-s|} \leq \frac{M-m}{\delta}$$

This follows from the property that for a convex function, the slope of the chord is non-decreasing. By comparing the slope on $[s, t]$ with the slopes involving the boundary points a and b, we obtain the linear growth bound.

Claim 3.3.4
f has linear growth in the neighborhood of $\mathbf{x}_0$.

Let $\mathbf{x} \in B_{r/2}(\mathbf{x}_0)$. Define $\varphi(t) = f(\mathbf{x}_0 + t\mathbf{u})$ where $\mathbf{u} = \dfrac{\mathbf{x} - \mathbf{x}_0}{\|\mathbf{x} - \mathbf{x}_0\|}$. $\varphi(t)$ is convex. Applying Claim 2 with $\delta = r/4$ and $t = \|\mathbf{x} - \mathbf{x}_0\|, s = 0$:

$$\frac{|f(\mathbf{x}) - f(\mathbf{x}_0)|}{\|\mathbf{x} - \mathbf{x}_0\|} \leq \frac{M-m}{r/4} = 4\left(\frac{M-m}{r}\right)$$

Claim 3.3.5
f is continuous at $\mathbf{x}_0$.

From Claim 3, $|f(\mathbf{x}) - f(\mathbf{x}_0)| \leq K\|\mathbf{x} - \mathbf{x}_0\|$ where $K = \dfrac{4(M-m)}{r}$. As $\mathbf{x} \to \mathbf{x}_0$, $\|\mathbf{x} - \mathbf{x}_0\| \to 0$, which implies $f(\mathbf{x}) \to f(\mathbf{x}_0)$. □

Theorem 3.4 (Supporting Hyperplane)
Let $C \subseteq \mathbb{R}^n$ be a non-empty convex set. Suppose $\mathbf{x}_0 \in \partial C$ (the boundary of C). Then there exists a non-zero vector $\mathbf{a} \in \mathbb{R}^n$ such that the hyperplane $H = \{\mathbf{x} : \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{x}_0\}$ supports C at $\mathbf{x}_0$. Specifically, the entire set C lies on one side of the hyperplane:

$$\mathbf{a}^T\mathbf{x} \geq \mathbf{a}^T\mathbf{x}_0 \quad \forall \mathbf{x} \in C$$

Proof. We consider two cases based on the interior of C.

Case 1: int$(C) \neq \emptyset$.
Since $\mathbf{x}_0$ lies on the boundary ∂C, it does not belong to the interior, i.e., $\mathbf{x}_0 \notin \text{int}(C)$. The set $\text{int}(C)$ is open and convex. The singleton set $\{\mathbf{x}_0\}$ is convex and compact. Since these two sets are disjoint, by the **Separating Hyperplane Theorem**, there exists a non-zero vector $\mathbf{a} \in \mathbb{R}^n$ that separates them. Thus, for all $\mathbf{y} \in \text{int}(C)$:

$$\mathbf{a}^T\mathbf{y} \geq \mathbf{a}^T\mathbf{x}_0$$

(Note: The direction of the inequality depends on the choice of $\mathbf{a}$, here we assume the direction pointing into the set). Since any point $\mathbf{x} \in C$ can be approached by a sequence of points in $\text{int}(C)$ (as $\overline{\text{int}(C)} = \overline{C}$ for convex sets with non-empty interior), and the inner product is continuous, the inequality holds for all $\mathbf{x} \in C$:

$$\mathbf{a}^T\mathbf{x} \geq \mathbf{a}^T\mathbf{x}_0$$

Case 2: $\mathbf{int}(C) = \emptyset$.

If the interior is empty, the convex set C must lie in an affine subspace of dimension $k < n$ (its relative interior is non-empty, but its full interior is empty). We can choose a vector $\mathbf{a}$ that is orthogonal to this affine subspace (treating $\mathbf{x}_0$ as the origin relative to the subspace). For such a vector, $\mathbf{a}^T(\mathbf{x} - \mathbf{x}_0) = 0$ for all $\mathbf{x} \in C$, which implies:

$$\mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{x}_0$$

This trivially satisfies the condition $\mathbf{a}^T\mathbf{x} \geq \mathbf{a}^T\mathbf{x}_0$. $\square$

Theorem 3.5 (Affine Majorization / Existence of Subgradient)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. For any point $\mathbf{x}_0$ where f is finite, the subdifferential set $\partial f(\mathbf{x}_0)$ is non-empty. That is, there exists at least one subgradient $\mathbf{g}$ such that the affine function $h(\mathbf{x}) = f(\mathbf{x}_0) + \mathbf{g}^T(\mathbf{x} - \mathbf{x}_0)$ supports f at $\mathbf{x}_0$.

Proof. Consider the point $(\mathbf{x}_0, f(\mathbf{x}_0)) \in \mathbb{R}^{n+1}$. This point lies on the boundary of the epigraph, $\text{epi}(f)$. Since f is convex, $\text{epi}(f)$ is a convex set. By the **Supporting Hyperplane Theorem**, there exists a non-zero normal vector $(\mathbf{a}, b) \in \mathbb{R}^n \times \mathbb{R}$ (where $(\mathbf{a}, b) \neq (\mathbf{0}, 0)$) such that the hyperplane passing through $(\mathbf{x}_0, f(\mathbf{x}_0))$ supports the epigraph. Thus, for all $(\mathbf{x}, t) \in \text{epi}(f)$, we have:

$$\begin{aligned} \begin{bmatrix} \mathbf{a}^T & b \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ t \end{bmatrix} &\leq \begin{bmatrix} \mathbf{a}^T & b \end{bmatrix} \begin{bmatrix} \mathbf{x}_0 \\ f(\mathbf{x}_0) \end{bmatrix} \\ \mathbf{a}^T\mathbf{x} + bt &\leq \mathbf{a}^T\mathbf{x}_0 + bf(\mathbf{x}_0) \end{aligned} \tag{1}$$

Step 1: Show $b \leq 0$.

Fix $\mathbf{x} = \mathbf{x}_0$. For any $t > f(\mathbf{x}_0)$, the point $(\mathbf{x}_0, t)$ is in $\text{epi}(f)$. Substituting into (1):

$$\mathbf{a}^T\mathbf{x}_0 + bt \leq \mathbf{a}^T\mathbf{x}_0 + bf(\mathbf{x}_0) \implies b(t - f(\mathbf{x}_0)) \leq 0$$

Since $t - f(\mathbf{x}_0) > 0$, we must have $b \leq 0$.

Step 2: Show $b \neq 0$.

Suppose for contradiction that $b = 0$. Then $\mathbf{a} \neq \mathbf{0}$ (since the normal vector is non-zero). The inequality (1) becomes:

$$\mathbf{a}^T\mathbf{x} \leq \mathbf{a}^T\mathbf{x}_0 \quad \forall \mathbf{x} \in \text{dom}(f)$$

If $\text{dom}(f) = \mathbb{R}^n$, this implies that a linear function is bounded above by a constant, which is only possible if $\mathbf{a} = \mathbf{0}$. This creates a contradiction (since $(\mathbf{a}, b) \neq \mathbf{0}$). Thus, $b < 0$.

Step 3: Construct the subgradient.
Since $b < 0$, we can divide inequality (1) by b (reversing the inequality sign):

$$\frac{\mathbf{a}^T}{b}\mathbf{x} + t \geq \frac{\mathbf{a}^T}{b}\mathbf{x}_0 + f(\mathbf{x}_0)$$

Let $\mathbf{g} = -\frac{1}{b}\mathbf{a}$. Then for any $\mathbf{x}$, taking $t = f(\mathbf{x})$ (which is in the epigraph):

$$-\mathbf{g}^T\mathbf{x} + f(\mathbf{x}) \geq -\mathbf{g}^T\mathbf{x}_0 + f(\mathbf{x}_0)$$

Rearranging terms yields the subgradient definition:

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + \mathbf{g}^T(\mathbf{x} - \mathbf{x}_0)$$

Thus, $\mathbf{g} \in \partial f(\mathbf{x}_0)$. $\square$