Indian Institute of Technology Kanpur
CS773 Online Learning and Optimization

Scribe: Sharbatanu Chatterjee, Kalpant Pathak
Instructor: Purushottam Kar
Date: January 10, 2017

LECTURE
2

Convex Analysis

1 Introduction

This lecture discusses a few mathematical preliminaries necessary to study convex analysis. Definitions and properties (with proofs wherever necessary) have been presented below.

2 Vector spaces

We begin with defining the ubiquitous concept of a vector space and several properties associated with them, as well as mathematical operations defined on them. The concept of a vector space is, in general, defined over an arbitrary field of *scalars*. However, we shall restrict ourselves to real vector spaces, i.e. those defined over the field of reals.

Definition 2.1 (Vector Space)**.** A vector space V defined over a field of scalars $\mathbb{F}$ is a set of mathematical objects called *vectors* along with two operations namely vector addition $V \times V \to V$ (often denoted $+$) and scalar multiplication $\mathbb{F} \times V \to V$ (often denoted $\cdot$) such that $(V, +)$ forms an abelian additive group and $(\mathbb{F}, \cdot)$ constitutes a multiplicative *action* over V. These properties are described in detail below.

Additive Abelian Group: For any $\mathbf{u}, \mathbf{v} \in V$, we have

1. *Closure under Vector Addition*: $\mathbf{u} + \mathbf{v} \in V$
2. *Commutativity*: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
3. *Associativity*: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
4. *Additive Identity*: $\exists\ \mathbf{0} \in V$ such that $\mathbf{v} + \mathbf{0} = \mathbf{v}$
5. *Additive Inverse*: For every $\mathbf{v} \in V$, $\exists\ \mathbf{v}' \in V$ such that $\mathbf{v} + \mathbf{v}' = \mathbf{0}$

Multiplicative Group Action: For any $\mathbf{u}, \mathbf{v} \in V$ and $a, b \in \mathbb{F}$, we have

1. *Closure under Scalar Mutiplication*: $a \cdot \mathbf{v} \in V$
2. *Distributivity over Vector Addition*: $a \cdot (\mathbf{u} + \mathbf{v}) = a \cdot \mathbf{u} + a \cdot \mathbf{v}$
3. *Distributivity over Scalar Addition*: $(a \oplus b) \cdot \mathbf{v} = a \cdot \mathbf{v} + b \cdot \mathbf{v}$
4. *Distributivity over Scalar Multiplication*: $(a \otimes b) \cdot \mathbf{v} = a \cdot (b \cdot \mathbf{v})$
5. *Normalization over Scalar Multiplication*: $1 \cdot \mathbf{v} = \mathbf{v}$

In the above properties, $\oplus, \otimes$ respectively denote the (scalar) addition and multiplication operations defined over the field $\mathbb{F}$ and $0, 1$ denote respectively the additive and multiplicative identities thereof.

Example 2.1. Some examples of vector spaces include:

1. $\mathbb{R}^n$, which is the set of all n-dimensional vectors with real components.
2. $B(\mathcal{X})$, which is the set of bounded functions over the domain $\mathcal{X}$.
3. $\mathcal{C}^1(\mathcal{X})$, which is the set of once differentiable functions over the domain $\mathcal{X}$.
4. $\mathcal{C}^\infty(\mathcal{X})$ which is the set of functions over the domain $\mathcal{X}$ which are infinitely differentiable (smooth functions).

Exercise 2.1. Prove that any two additive identity vectors in a vector space are indiscernible. More specifically, show that if there exist two identity vectors $\mathbf{0}_1, \mathbf{0}_2$, then $\mathbf{0}_1 = \mathbf{0}_2$.

Exercise 2.2. Show that for any vector $\mathbf{v} \in V$, we have $0 \cdot \mathbf{v} = \mathbf{0}$, where 0 is the multiplicative identity of the scalar field $\mathbb{F}$ and $\mathbf{0}$ is the additive identity of the vector space V.

3 Inner Product

The previous section introduced us to the concept of vector spaces with simple operations of vector addition and scalar multiplication. We now describe more some interesting operations on vector spaces that endow it with beautiful structure and topology. The concept of an inner product is a natural extension of the standard dot product.

Definition 2.2 (Inner Product)**.** An inner product is a real valued (in general $\mathbb{F}$-valued) bivariate function on vector spaces $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ having the following properties.

Definition 2.3 (Inner Product Space)**.** A normed space, $(V, \|.\|)$, is a vector space over which a norm has been established.

Remark 2.1. An inner product space that is complete with respect to the norm induced by the inner product (see below for definition of induced norms) is called a Hilbert space.

1. *Linearity w.r.t Vector Addition*: For all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in V$, $\langle \mathbf{u}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$
2. *Linearity w.r.t Scalar Multiplication*: For all $a \in \mathbb{R}$, $\langle a \cdot \mathbf{u}, \mathbf{v} \rangle = a \cdot \langle \mathbf{u}, \mathbf{v} \rangle$
3. *Commutativity*: $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$
4. *Positive Definiteness*: $\forall \mathbf{u} \neq \mathbf{0}, \langle \mathbf{u}, \mathbf{u} \rangle > 0$

Example 2.2. Some examples of inner products include:

1. $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^d x_i y_i$ for vectors in the d-dimensional space.[1]
2. *(Weighted inner products)* For any vector $\mathbf{w} > \mathbf{0}$[2], we define $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{w}} = \sum_{i=1}^d w_i x_i y_i$

[1] This is the notion of *dot product* we generally have for vectors, also denoted frequently as $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$

[2] For two d-dimensional vectors $\mathbf{u}, \mathbf{v}$, the notation $\mathbf{u} > \mathbf{v}$ implies a coordinate-wise relation i.e. for all $i \in [d]$, $u_i > v_i$. We define the relation $\mathbf{u} \geq \mathbf{v}$ similarly.

3. *(Inner product induced by a matrix)* For any real positive definite (PD) matrix[3] A, $(A \succ 0)$, we define $\langle \mathbf{x}, \mathbf{y} \rangle_A := \mathbf{x}^\top A \mathbf{y}$

4. *(Inner product over function spaces)* For any two real valued square summable functions over a domain $\mathcal{X}$, $f, g : \mathcal{X} \to \mathbb{R}$, we can define an inner product as $\langle f, g \rangle = \int_{\mathcal{X}} f(x)g(x)dx$

Exercise 2.3. Show that the condition $A \succ 0$ is essential for $\langle \mathbf{x}, \mathbf{y} \rangle_A := \mathbf{x}^\top A \mathbf{y}$ to be a valid inner product.

Exercise 2.4. Show that the condition $\mathbf{w} > \mathbf{0}$ is essential for $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{w}} = \sum_{i=1}^d w_i x_i y_i$ to be a valid inner product.

Exercise 2.5. Show that $\langle \mathbf{v}, \mathbf{0} \rangle = 0$ for all $\mathbf{v} \in V$ where $\mathbf{0}$ is the (vector) additive identity of V.

4 Norm

Inner products gave us an idea about the interaction of two vectors. This section introduces us to the concept of *norm* which corresponds to abstract notions of *lengths* of vectors in a vector space as well as notions of *distance* between two vectors.

Definition 2.4 (Norm)**.** A norm is a real valued univariate function $\|\cdot\| = V \to \mathbb{R}$ having the following properties.

Definition 2.5 (Normed Space)**.** A normed space, $(V, \|.\|)$, is a vector space over which a norm has been established.

Remark 2.2. A *complete* normed space (i.e. one in which all sequences that are Cauchy with respect to that norm converge) is called a Banach space.

1. *Separation of the discernibles* If $\|\mathbf{v}\| = 0$, then $\mathbf{v} = \mathbf{0}$ [4]

2. *Positive Homogeneity*: For any scalar $a \in \mathbb{R}$, $\|a \cdot \mathbf{v}\| = |a| \cdot \|\mathbf{v}\|$

3. *Triangle Inequality/Subadditivity*: $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.[5]

Example 2.3. Some examples of commonly used norms are:

1. (ℓ_p norms, $1 \leq p < \infty$) $\|\mathbf{x}\|_p := \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$. The ℓ_2 norm is often referred to as the *Euclidean norm.* The ℓ_1 norm is often referred to as the *Manhattan norm.*

2. l_∞ norm, defined as $\|\mathbf{x}\|_\infty := \max_i |x_i|$

3. *(Norms induced by an inner product)* Corresponding to every inner product defined over a vector space, we can define a norm as $\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$. Such a norm is said to be *induced* by the corresponding inner product.

4. *(Norms induced by a matrix)* defined as $\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^\top A \mathbf{x}}$. These are often called Mahalanobis norms.

Exercise 2.6. Show that all norms are symmetric i.e. $\|\mathbf{u} - \mathbf{v}\| = \|\mathbf{v} - \mathbf{u}\|$ for all $\mathbf{u}, \mathbf{v} \in V$.

---

[3] A PD matrix is a square symmetric matrix with all eigenvalues > 0

[4] The concept behind this property is that if two vectors are distinct from each other, the "distance" between them should be non-zero. If $\mathbf{u} \neq \mathbf{v}$ are two distinct vectors, then $\|\mathbf{u} - \mathbf{v}\| \neq 0$.

[5] The intuition behind this is that a detour should increase the "distance" travelled.

Exercise 2.7. Show that all norms are non-negative valued i.e. $\|\mathbf{v}\| \geq 0$ for all $\mathbf{v} \in V$. This is the so-called *Positivity* property of norms. As a corollary, show that all norms satisfy the so-called *Positive Separation* property, i.e. $\|\mathbf{v}\| > 0$ for all $\mathbf{v} \neq \mathbf{0}$. [6]

Exercise 2.8. Show that all norms induced by inner products are valid norms.

Exercise 2.9. Show that norms are induced only by those matrices that are positive definite, i.e. all their eigenvalues are strictly positive.

Exercise 2.10. A very widely used "norm" is the sparsity or ℓ_0 norm defined as $\|\mathbf{x}\|_0 = |i : x_i \neq 0|$. However, this is not a valid norm. Which properties of norms does this violate?

5 Dual Spaces and Dual Norms

We first introduce the notion of the dual of a normed space.

Definition 2.6 (Linear Functional)**.** A real valued univariate function over a normed space $f : V \to \mathbb{R}$ is said to be a *linear functional* if it satisfies the following properties:

- $f(\mathbf{u}) + f(\mathbf{v}) = f(\mathbf{u} + \mathbf{v})$, $\forall \mathbf{u}, \mathbf{v} \in V$.
- $f(a \cdot \mathbf{u}) = a \cdot f(\mathbf{u})$, $\forall \mathbf{u} \in V$ and $a \in \mathbb{R}$

A linear functional f is said to be *bounded* if $\sup_{\|\mathbf{x}\| \leq 1} \ f(\mathbf{x}) < \infty$ i.e. it takes bounded values inside the unit ball.

Definition 2.7 (Dual Space)**.** The vector space of all bounded linear functionals over a vector space V, which we shall denote as $\mathcal{F}_{\text{lin}}(V)$, is referred to as the *dual* space of V. Another popular notation for the dual space is V^*.

In the exercises, we will attempt to visualize what $\mathcal{F}_{\text{lin}}$ might look like when the vector space is simple. We now define the dual norm of a vector space as follows:

Definition 2.8 (Dual Norm)**.** Given a linear functional space $\mathcal{F}_{\text{lin}}$, associated with a normed vector space $(V, \|\cdot\|)$, the dual norm defined on $\mathcal{F}_{\text{lin}}$ is given by

$$\|f\|_* := \sup \{f(x), \|x\| \leq 1\}$$

Example 2.4. Examples of dual norms are:

1. $\|.\|_2$ is the dual norm of itself.
2. $\|.\|_1$ is the dual norm of $\|.\|_\infty$
3. The dual norm of $\|.\|_p$ is $\|.\|_q$ such that $1/p + 1/q = 1$

Exercise 2.11. For any vector spaces V, consider an additive operation $\boxplus$ defined on $\mathcal{F}_{\text{lin}}(V)$ as follows: for any $f, g \in \mathcal{F}_{\text{lin}}$, define $f \boxplus gh$ so that for any $\mathbf{v} \in V, (f \boxplus g)(\mathbf{v}) = f(\mathbf{v}) + g(\mathbf{v})$.

1. Show that for any $f, g \in \mathcal{F}_{\text{lin}}$, $f \boxplus g \in \mathcal{F}_{\text{lin}}$.
2. Show that the dual space $(\mathcal{F}_{\text{lin}}(V), \boxplus)$ is a valid vector space over the scalar field of reals.

[6] Norms for which this property does not hold i.e. $\|\mathbf{u}\| = 0$ for some $\mathbf{u} \neq \mathbf{0}$ are called *degenerate norms* or *semi-norms*. These violate the property of separation of the discernibles. By definition, norms are non-degenerate.

Exercise 2.12. Show that for any normed space $(V, \|\cdot\|)$, the dual norm $\|\cdot\|_*$ defined on the dual space is a valid norm.

Exercise 2.13. Show that a bounded linear functional takes a bounded value on every vector with a bounded norm i.e. if f is a bounded linear functional then $f(\mathbf{x}) < \infty$ for all $\|\mathbf{x}\| < \infty$.

Exercise 2.14. Show that when the vector space in question is an inner product space, more specifically a finite dimensional Euclidean space with the standard inner product, the dual space is isometrically isomorphic to it. More specifically, if the normed space we are working with is $(\mathbb{R}^n, \|\cdot\|_2)$ (recall that the ℓ_2 norm is induced by the standard inner product), then

1. Show that $\mathcal{F}_{\text{lin}}(\mathbb{R}^n)$ is isomorphic to $\mathbb{R}^n$.
2. Show that the norm dual to the ℓ_2 norm is the ℓ_2 norm itself.

The solution to a part of this exercise is given in Appendix B but do try it on your own before looking at the solution.

6 Cauchy-Schwartz Inequality

The Cauchy-Schwartz inequality gives a relation between the inner product of two vectors and the product of their ℓ_2 or Euclidean norms.

Theorem 2.1. *(Cauchy-Schwartz Inequality)* $|\langle \mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$

Proof. We will consider two separate cases for the proof.

1. *Case 1* ($\|\mathbf{u}\| = 0$ or $\|\mathbf{v}\| = 0$) Since norms are non-degenerate, this implies that either $\mathbf{u} = \mathbf{0}$ or $\mathbf{v} = \mathbf{0}$. Thus, both the left and the right hand sides of the inequality are zero and the inequality holds.
2. *Case 2* ($\mathbf{u} \neq 0, \mathbf{v} \neq 0$) In this case we can assume, w.l.o.g. [7], that $\|\mathbf{u}\| = 1, \|\mathbf{v}\| = 1$. Now we only need to prove that
$$|\langle \mathbf{u}, \mathbf{v}\rangle| \leq 1$$
Consider the vector $\mathbf{u}_\perp = \mathbf{u} - \langle \mathbf{u}, \mathbf{v}\rangle \mathbf{v}$. It is easy to see that $\langle \mathbf{u}_\perp, \mathbf{v}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle \langle \mathbf{v}, \mathbf{v}\rangle = \langle \mathbf{u}, \mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle \|\mathbf{v}\|^2 = \langle \mathbf{u}, \mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle = 0$. Moreover, it is easy to see that $\mathbf{u} - \mathbf{u}_\perp$ is aligned to $\mathbf{v}$. Thus, $\mathbf{u}_\perp$ is the entire component of $\mathbf{u}$ orthogonal to $\mathbf{v}$.

   Since we have already shown in an exercise that induced norms are valid norms, by the positivity property of norms, we have $\|\mathbf{u}_\perp\|_2^2 \geq 0$ which gives us
$$0 \leq \langle \mathbf{u} - \langle \mathbf{u}, \mathbf{v}\rangle \mathbf{v}, \mathbf{u} - \langle \mathbf{u}, \mathbf{v}\rangle \mathbf{v}\rangle = \|\mathbf{u}\|^2 + \|\langle \mathbf{u}, \mathbf{v}\rangle \mathbf{v}\|^2 - 2\langle \mathbf{u}, \langle \mathbf{u}, \mathbf{v}\rangle \mathbf{v}\rangle = 1 - \langle \mathbf{u}, \mathbf{v}\rangle^2,$$

which concludes the proof. □

Exercise 2.15 (Generalized Cauchy-Schwartz Inequality). Show that for any real valued linear functional f on a normed space $(V, \|\cdot\|)$, we have $|f(\mathbf{x})| \leq \|f\|_* \|\mathbf{x}\|$ for any $\mathbf{x} \in V$ where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

Exercise 2.16. Show that the Cauchy-Schwartz inequality also holds for semi-norms. The solution to this exercise is given in Appendix A but do try it on your own before looking at the solution.

[7] Without loss of generality

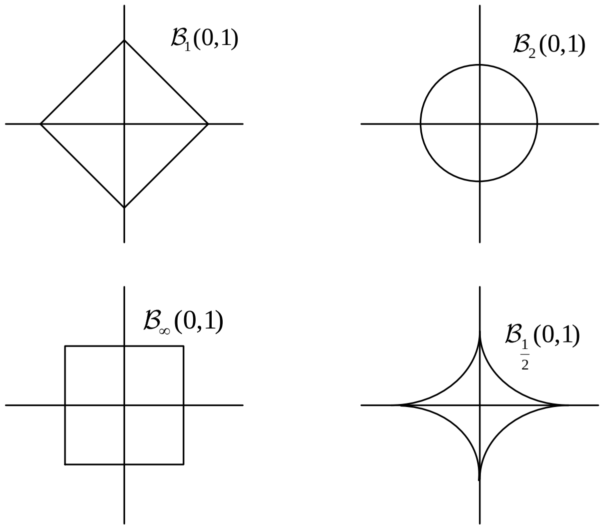

Figure 1: Graphical representation of two-dimensional norm balls associated with different norms. Note that $\ell_{1/2}$ (and in general ℓ_p for any $p < 1$) is not a norm although it is often referred to as one.

7 Hyperplanes and Halfspaces

The next mathematical object we shall deal with are *hyperplanes* defined as follows.

Definition 2.9 (Hyperplane)**.** For every $f \in \mathcal{F}_{\text{lin}}$ (i.e. a linear functional), we define a corresponding hyperplane, $\mathcal{H}_f$ as

$$\mathcal{H}_f := \{\mathbf{v} : f(\mathbf{v}) = 0\}$$

As we can see, in vectors over $\mathbb{R}^2$, all straight lines are hyperplanes. In vectors over the Euclidean space $\mathbb{R}^3$, all planes are hyperplanes, given by $\mathcal{H}_{\mathbf{a}} = \left\{\mathbf{v} : \mathbf{a}^\top \mathbf{v} = 0\right\}$

Corresponding to every hyperplane, there exists two *halfspaces* defined as the set of vectors on either "side" of the hyperplane.

Definition 2.10 (Halfspace)**.** For every $f \in \mathcal{F}_{\text{lin}}$, we define the halfspace $\mathcal{E}_f$ as

$$\mathcal{E}_f = \{\mathbf{v} : f(\mathbf{v}) \geq 0\}.$$

Halfspaces may be defined using either a strict ($>$) or a non-strict ($\geq$) inequality. For example, corresponding to linear form $\mathbf{v} \mapsto \mathbf{a}^\top \mathbf{v}$, we have the halfspace $\mathcal{E}_{\mathbf{a}} = \left\{\mathbf{v} : \mathbf{a}^\top \mathbf{v} \geq 0\right\}$.

8 Balls and Ellipsoids

Definition 2.11 (Ball)**.** A ball, corresponding to a norm, is defined as $\mathcal{B}_{\|\cdot\|}(\mathbf{v}, r) := \{\mathbf{x} : \|\mathbf{x} - \mathbf{v}\| \leq r\}$. The vector $\mathbf{v}$ is called the *center* of the ball and the scalar r is called the radius of the ball. The notation $\mathcal{B}_{\|\cdot\|}(r)$ is often used to refer to $\mathcal{B}_{\|\cdot\|}(\mathbf{0}, r)$.

Example 2.5. $\mathcal{B}_2(\mathbf{0}, r) = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq r\}$

Definition 2.12 (Ellipsoid)**.** An ellipsoid, similar to a ball, is defined as $\mathcal{E}_A(\mathbf{v}, r) := \{\mathbf{x} : \|\mathbf{x} - \mathbf{v}\|_A \leq r\}$. An ellipsoid is induced by a positive definite matrix A.

Remark 2.3. All ellipsoids are balls as they are balls corresponding to norms induced by positive definite matrices, but not all balls are ellipsoids.

9 Convex Sets

To introduce convex sets, we first need to study various types of combinations of vectors. Combinations are important in our analysis, as we shall often define 'sets' generated by, or closed under, combinations of vectors.

1. *Linear*: A linear combination of two vectors $\mathbf{u}, \mathbf{v} \in V$ is a vector of the form $\lambda \cdot \mathbf{u} + \mu \cdot \mathbf{v}$, where the combination coefficients λ and μ are real scalars. A set of vectors that is closed under linear combinations is called a linear set or more commonly a linear subspace.

2. *Affine*: An affine combination of the vectors $\mathbf{u}$ and $\mathbf{v}$ is a linear combination where the coefficients add up to one i.e. of the form $\lambda \cdot \mathbf{u} + \mu \cdot \mathbf{v}$, where λ, μ are such that $\lambda + \mu = 1$. A set of vectors closed under affine combinations is called an affine set or an affine subspace.

3. *Conic*: A conic combination of the vectors $\mathbf{u}$ and $\mathbf{v}$ is a linear combination where the coefficients are non-negative i.e of the form $\lambda \cdot \mathbf{u} + \mu \cdot \mathbf{v}$, for any $\lambda \geq 0, \mu \geq 0$. A set of vectors closed under conic combinations is called a conic set or simply a cone.

4. *Convex*: A convex combination of the vectors $\mathbf{u}$ and $\mathbf{v}$ is a linear combination where the coefficients are non-negative and add up to one i.e. of the form $\lambda \cdot \mathbf{u} + (1 - \lambda) \cdot \mathbf{v}$, for any $\lambda \in [0, 1]$. A set of vectors closed under convex combinations is called a convex set.

Definition 2.13 (Convex Set)**.** A set that is closed under all convex combinations is called a convex set. That is to say a set $\mathcal{C} \subset V$ is called convex if for all $\mathbf{u}, \mathbf{v} \in \mathcal{C}$ and $\lambda \in [0, 1]$, we have $\lambda \cdot \mathbf{u} + (1 - \lambda) \cdot \mathbf{v} \in \mathcal{C}$.

In Euclidean spaces, the above definition of convex sets can be interpreted as requiring a convex set to contain all line segments joining two points inside that set. This is because the set of all convex combinations of two points in a Euclidean space is simply the line segment joining those two points. An illustrative example of a convex and a non-convex set is given in Figure 2.

Remark 2.4. Given a set $S \subset V$ of points (finite/infinite, discrete/non-discrete) in a vector space, the smallest linear subspace containing all these points is often called the linear subspace *generated* by S. Similarly the smallest affine/conic set containing S is called the affine/conic set generated by S. The smallest convex set containing S has a special name – it is called the *convex hull* of S.

Exercise 2.17. Show that the intersection of two convex sets is always convex. Also argue using a counter example that the union of two convex sets is not necessarily convex. Under what conditions is the union of two convex sets also convex?

Exercise 2.18. Show the following

1. All linear subspaces are vector spaces.

2. All hyperplanes are affine sets.

3. All halfspaces are convex sets.

4. The set of all positive semi-definite matrices is conic.

5. All balls generated by norms are convex sets.

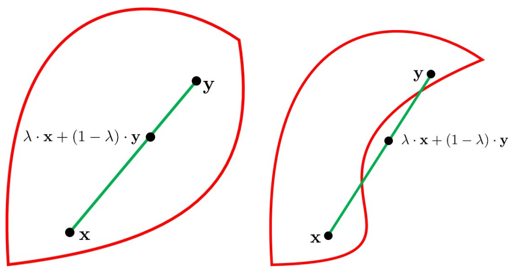

Figure 2: A convex and a non-convex set.

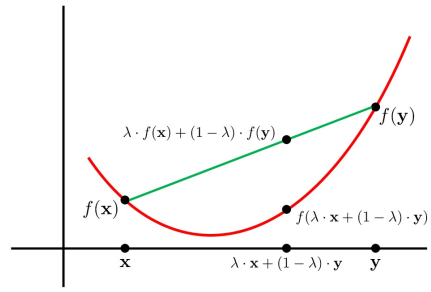

Figure 3: A real convex function.

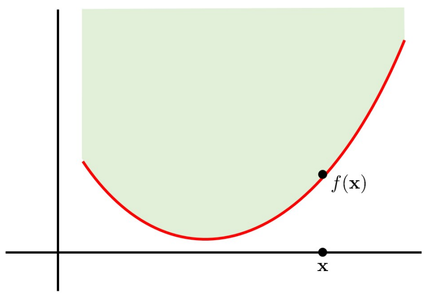

Figure 4: An epigraph-convex function.

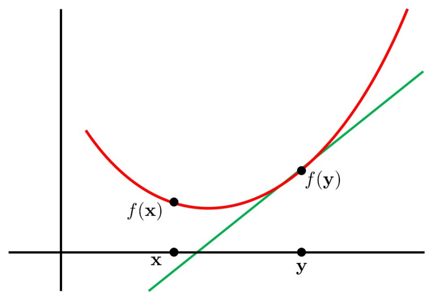

Figure 5: A differentiable convex function.

10 Convex functions

We finally arrive at defining convex functions. We will look at several definitions of convex functions, some more general, some more specific, some that are more fundamental and others that are easy to use.

Definition 2.14 (Convex Function)**.** A function $f : V \to \mathbb{R}$, is called convex if $\forall\ \mathbf{u}, \mathbf{v} \in V, \lambda \in [0, 1]$,

$$f(\lambda \cdot \mathbf{u} + (1-\lambda) \cdot \mathbf{v}) \leq \lambda f(\mathbf{u}) + (1-\lambda) f(\mathbf{v}).$$

See Figure 3 for an illustration.

Some texts use a slightly different notion of convexity based on the epigraph of the function. We state that definition below for the case of functions on Euclidean spaces for sake of simplicity.

Definition 2.15 (Epigraph-convex Function)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ with domain $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ (assume that the domain is convex) is said to be epigraph convex if the epigraph of the function, defined as the set $\mathrm{epi}(f) = \left\{(\mathbf{x}, y) \in \mathbb{R}^{n+1}, \mathbf{x} \in \mathrm{dom}(f), y \geq f(\mathbf{x})\right\}$ is convex. See Figure 4 for an illustration.

These fundamental definitions are equivalent but cumbersome to work with. A weaker but more workable definition is that of *mid-point convexity*:

Definition 2.16 (Midpoint-convex function)**.** A function is called mid-point convex if $\forall\ \mathbf{u}, \mathbf{v} \in V$,

$$f\left(\frac{\mathbf{u}+\mathbf{v}}{2}\right) \leq \frac{f(\mathbf{u}) + f(\mathbf{v})}{2}.$$

Theorem 2.2. A continuous function is convex iff it is mid-point convex. (For a proof, originally by Jensen in 1905, refer to Sra et al. (2014))

A still more convenient definition of convex functions is presented for differentiable functions. However, we need to define the notion of the differential of a function on a normed space.

10.1 Derivatives of functions

Definition 2.17 (Fréchet derivatives)**.** A real-valued function $f : V \to \mathbb{R}$ defined over a normed space is said to possess a Fréchet derivative $g \in \mathcal{F}_{\text{lin}}$ at a point $\mathbf{x} \in V$ if the following limiting behavior exists

$$\lim_{\mathbf{h} \to \mathbf{0}} \frac{|f(\mathbf{x} + \mathbf{h}) - f(\mathbf{x}) - g(\mathbf{h})|}{\|h\|} = 0$$

The derivative g is often referred to as the *gradient* of f.

Lemma 2.3. A differentiable function is convex iff

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle ,$$

where ∇f is the gradient of the function f.

Proof. (Adapted from Boyd and Vandenberghe (2004)). Since f is continuous, we will use the midpoint convexity definition to prove the result.

(If) Consider $\mathbf{z} = \frac{\mathbf{x}+\mathbf{y}}{2}$. The gradient condition gives us

$$f(\mathbf{x}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle$$
$$f(\mathbf{y}) \geq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle$$

Adding the equations and dividing by two gives us the required result

$$\frac{f(\mathbf{x}) + f(\mathbf{y})}{2} \geq f(\mathbf{z}) = f\left(\frac{\mathbf{x} + \mathbf{y}}{2}\right)$$

(Only if) For sake of simplicity, assume that f is defined and differentiable over the entire vector space. Then for $\lambda \in (0, 1]$, we have, by the convexity property

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

which, upon rearranging, gives us

$$f(\mathbf{x}) \geq f(\mathbf{y}) + \frac{f(\mathbf{y} + \lambda \cdot (\mathbf{x} - \mathbf{y})) - f(\mathbf{y})}{\lambda}.$$

Now, we have

$$\begin{aligned} \frac{f(\mathbf{y} + \lambda \cdot (\mathbf{x} - \mathbf{y})) - f(\mathbf{y})}{\lambda} &= \frac{f(\mathbf{y} + \lambda \cdot (\mathbf{x} - \mathbf{y})) - f(\mathbf{y})}{\lambda} \\ &= \underbrace{\frac{f(\mathbf{y} + \lambda \cdot (\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \lambda \cdot (\mathbf{x} - \mathbf{y}) \rangle}{\lambda}}_{(A)} + \langle \nabla f(\mathbf{y}), (\mathbf{x} - \mathbf{y}) \rangle \end{aligned}$$

Now we have

$$\begin{aligned} (A) &= \frac{f(\mathbf{y} + \lambda \cdot (\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \lambda \cdot (\mathbf{x} - \mathbf{y}) \rangle}{\lambda} \\ &= \frac{f(\mathbf{y} + \lambda \cdot (\mathbf{x} - \mathbf{y})) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \lambda \cdot (\mathbf{x} - \mathbf{y}) \rangle}{\lambda \cdot \|\mathbf{x} - \mathbf{y}\|} \cdot \|\mathbf{x} - \mathbf{y}\| \end{aligned}$$

Taking $\lambda \to 0$ and using Definition 2.17 gives us $(A) \to 0$ which proves the result. □

Remark 2.5. A differentiable function $f : V \to \mathbb{R}$ is said to be *invex* if there exists a combinator function $g : V \times V \to V$ such that for all $\mathbf{x}, \mathbf{y} \in V$, we have $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), g(\mathbf{x}, \mathbf{y}) \rangle$. Convex differentiable functions are invex with the combinator function $g(\mathbf{x}, \mathbf{y}) = \mathbf{x} - \mathbf{y}$.

Remark 2.6. For twice differentible functions, a still more convenient definition exists: a twice differentiable function is convex iff its Hessian is positive semi definite everywhere i.e. $\nabla^2 f(\mathbf{x}) \succeq 0$ for all $\mathbf{x} \in \mathrm{dom}(f)$.

Exercise 2.19. Prove that a function is convex iff it is epigraph convex.

Exercise 2.20. Prove that all norms are convex functions.

Exercise 2.21. Let $\mathcal{F} = \{f : V \to \mathbb{R}\}$ be a family of convex real-valued functions on a vector space. Prove that the function $g : \mathbf{x} \mapsto \sup_{f \in \mathcal{F}} f(\mathbf{x})$ is also a convex function. *Hint*: First try to prove the result in the case where the supremum is assured to be always achieved i.e. for all $\mathbf{x} \in \mathcal{F}$, there exists $f_{\mathbf{x}} \in \mathcal{F}$ such that $g(\mathbf{x}) = f_{\mathbf{x}}(\mathbf{x})$.

Definition 2.18 (Lipschitz Function)**.** A function $f : V \to \mathbb{R}$ on a normed space V, is said to be L-Lipschitz if

$$|f(\mathbf{u}) - f(\mathbf{v})| \leq L \left\|\mathbf{u} - \mathbf{v}\right\|, \; \forall \mathbf{u}, \mathbf{v} \in V$$

where L is a constant. In other words, a Lipschitz function can not change values too abruptly. For differentiable functions, this happens if the gradient is norm bounded.

Lemma 2.4. Let $f : V \to \mathbb{R}$ be a differentiable function such that $\|\nabla f(\mathbf{x})\| \leq L$. Then f is L-Lipschitz.

Proof. The mean value theorem states that if f is differentiable, then for any $\mathbf{u}, \mathbf{v}$, there exists a $\lambda \in [0, 1]$ such that

$$f(\mathbf{u}) - f(\mathbf{v}) = \langle \nabla f((\lambda \cdot \mathbf{u} + (1 - \lambda) \cdot \mathbf{v})), \mathbf{u} - \mathbf{v} \rangle .$$

Applying the Cauchy-Schwartz inequality and the fact that $\|\nabla f((\lambda \cdot \mathbf{u} + (1 - \lambda) \cdot \mathbf{v}))\| \leq L$ proves the result. □

11 Convex Projections

Consider a constrained convex optimization problem

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

where $f(\cdot)$ is a convex function and $\mathcal{C}$ is a closed convex set (in practive this set arises due to various convex and/or affine constraints). These problems arise in several machine learning and signal processing applications. A popular way solving these constrained optimization problems is the *projected gradient descent method* which involves taking steps opposite to the direction of the gradient and projecting back onto the constraint. The convex projection step is crucial to this procedure. We analyze some properties of convex projections now.

Definition 2.19 (Convex Projection)**.** Let $\mathcal{C} \subset V$ be a closed convex set and $\|\cdot\|$ be a (non-degenerate) norm on the vector space V. Then the convex projection of a vector $\mathbf{z} \in V$ onto the set $\mathcal{C}$ is defined as:

$$\Pi_{\mathcal{C}}(\mathbf{z}) := \arg\min_{\mathbf{x} \in \mathcal{C}} \|\mathbf{x} - \mathbf{z}\| \tag{1}$$

The problem of fining a convex projection is a convex optimization problem itself but in well behaved cases, (1) has closed form solutions or the solution can be obtained in polynomial time.

11.1 Properties of Convex Projections

In the following, we will look at projections with respect to the ℓ_2 norm.

11.1.1 Property I

If $\hat{\mathbf{z}}$ is the projection of $\mathbf{z} \in V$ onto a convex set $\mathcal{C} \subset V$, then

$$\langle \mathbf{x} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle \leq 0, \; \forall \mathbf{x} \in \mathcal{C} \tag{2}$$

in other words, the angle between the vectors $\mathbf{x} - \hat{\mathbf{z}}$ and $\mathbf{z} - \hat{\mathbf{z}}$ is always greater than 90°. Intuitively, the case when angle between vectors $\mathbf{x} - \hat{\mathbf{z}}$ and $\mathbf{z} - \hat{\mathbf{z}}$ becomes less than 90° is only possible when $\mathcal{C}$ is a non-convex set.

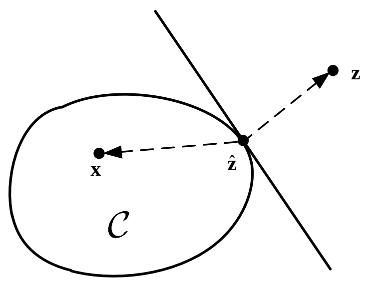

Figure 6: Graphical representation of projection property-I. The angle between the vectors $\mathbf{x} - \hat{\mathbf{z}}$ and $\mathbf{z} - \hat{\mathbf{z}}$ is less greater 90°.

Proof. (Bertsekas, 2010) The supporting hyperplane for the convex set $\mathcal{C}$ passing through vector $\hat{\mathbf{z}}$ is given as:

$$\langle \mathbf{w} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle = 0$$

We note that proving that the convex set is contained in the halfspace $\mathcal{C} \subseteq \mathcal{H} = \{\mathbf{w} | \, \langle \mathbf{w} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle \leq 0\}$ will establish the claimed result. Now, the projection $\hat{\mathbf{z}}$ is obtained by minimizing the following convex function:

$$f(\mathbf{x}) = \frac{1}{2} \left\| \mathbf{z} - \mathbf{x} \right\|_2^2 .$$

However, a point $\hat{\mathbf{z}}$ minimizes a function f over a convex set $\mathcal{C}$ if and only if

$$\langle \nabla f(\hat{\mathbf{z}}), \mathbf{x} - \hat{\mathbf{z}} \rangle \geq 0, \quad \forall \mathbf{x} \in \mathcal{C}$$

Since $\nabla f(\hat{\mathbf{z}}) = \hat{\mathbf{z}} - \mathbf{z}$, this condition becomes equivalent to (2). This completes the proof. □

11.1.2 Property II

If $\hat{\mathbf{z}}$ be the projection of $\mathbf{z} \in V$ onto a convex set $\mathcal{C} \subset V$, then

$$\left\| \mathbf{x} - \mathbf{z} \right\|_2 \geq \left\| \mathbf{x} - \hat{\mathbf{z}} \right\|_2 , \; \forall \mathbf{x} \in \mathcal{C} \tag{3}$$

i.e., $\hat{\mathbf{z}}$ is closer to all the points in $\mathcal{C}$ than $\mathbf{z}$.

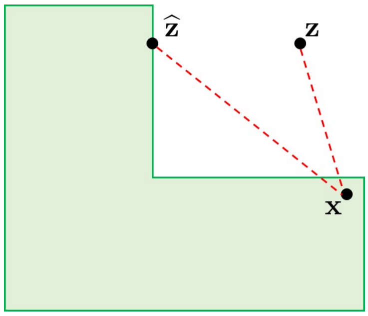

Figure 7: Projecting onto non-convex sets may cause us to move away from portions of the set

Proof. We have, for any $\mathbf{x} \in \mathcal{C}$,

$$\begin{aligned}
\|\mathbf{x} - \mathbf{z}\|_2^2 &= \|\mathbf{x} - \hat{\mathbf{z}} + \hat{\mathbf{z}} - \mathbf{z}\|_2^2 \\
&= \|\mathbf{x} - \hat{\mathbf{z}}\|_2^2 + \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2 - 2\underbrace{\langle \mathbf{x} - \hat{\mathbf{z}}, \mathbf{z} - \hat{\mathbf{z}} \rangle}_{\leq 0 \text{ (Property I)}} \\
&\geq \|\mathbf{x} - \hat{\mathbf{z}}\|_2^2 + \underbrace{\|\mathbf{z} - \hat{\mathbf{z}}\|^2}_{\geq 0} \\
&\geq \|\mathbf{x} - \hat{\mathbf{z}}\|_2^2 .
\end{aligned}$$

□

Property II of convex projections is immensely useful for convex optimizers since it shows that no matter where our optimum lies within the convex set and where are we projecting from, we will always move closer to the optimum.

Contrast this with the following example (see Figure 7) of a non-convex set where the projection takes us farther from certain parts of the set. If the optimum point lay in these parts, we were better off without projecting! However, for well behaved non-convex problems such as those that arise in sparse recovery, compressive sensing and matrix completion problems, a large variety of results have shown that non-convex projection are not badly behaved.

Exercise 2.22. Show that convex projections are idempotent i.e. $\Pi_{\mathcal{C}}(\mathbf{z}) = \mathbf{z}$ if $\mathbf{z} \in \mathcal{C}$.

Exercise 2.23. Devise an algorithm to project onto $\mathcal{C} = \mathcal{B}_\infty(\mathbf{0}, r)$, the ℓ_∞ ball of radius r.

Exercise 2.24. Devise an algorithm to project onto $\mathcal{C} = \mathcal{B}_2(\mathbf{0}, r)$ the ℓ_2 ball of radius r. Show that it has a closed form solution given as:

$$\Pi_{\mathcal{C}}(\mathbf{z}) = \begin{cases} \mathbf{z}, & \text{if } \mathbf{z} \in \mathcal{C} \\ \frac{\mathbf{z}}{\|\mathbf{z}\|} \cdot r, & \text{if } \mathbf{z} \notin \mathcal{C}. \end{cases}$$

Exercise 2.25. Devise an algorithm to project onto $\mathcal{C} = \mathcal{B}_1(\mathbf{0}, r)$, the ℓ_1 ball of radius r. *Hint*: Try to first project onto the simplex $\mathcal{C} = \{\mathbf{x} : \mathbf{x} \geq 0, \|\mathbf{x}\| \leq 1\}$. If you are unable to solve the problem, search online. Solutions to this problem have been rediscovered several times.

References

Dimitri P. Bertsekas. *Convex Optimization Theory*. Univ. Press, 2010.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Suvrit Sra, Avinava Dubey, and Ahmed Hefny. Scribed notes for Lecture 2, 10-801: Advanced Optimization and Convex functions, 2014. URL `http://www.cs.cmu.edu/~suvrit/teach/10801_lecture2.pdf`.

A The Cauchy-Schwartz Inequality for Semi-norms

To establish this, we need only analyze the case when $\|\mathbf{u}\| = 0$ which the following claim does.

Claim 2.5. If $\|\mathbf{u}\| = 0$ then $\langle \mathbf{v}, \mathbf{u} \rangle = 0$ for all $\mathbf{v} \in V$.

Proof. Consider any vector $\mathbf{v}$ and look at the vector $\mathbf{z} = \mathbf{v} - r \cdot \mathbf{u}$, where $r = t \cdot \langle \mathbf{u}, \mathbf{v} \rangle$ for some real value t (note that in the proof of the Cauchy-Schwartz inequality for norms, we had taken $r = \langle \mathbf{u}, \mathbf{v} \rangle$ i.e. $t = 1$). This gives us

$$\begin{aligned}
\|\mathbf{z}\|_2^2 &= \|\mathbf{v}\|_2^2 + r^2 \|\mathbf{u}\|_2^2 - 2r \cdot \langle \mathbf{u}, \mathbf{v} \rangle \\
&= \|\mathbf{v}\|_2^2 - 2r \cdot \langle \mathbf{u}, \mathbf{v} \rangle \\
&= \|\mathbf{v}\|_2^2 - 2t \cdot \langle \mathbf{u}, \mathbf{v} \rangle^2 .
\end{aligned}$$

Since, $\|\mathbf{z}\|_2 \geq 0$, this means $\langle \mathbf{u}, \mathbf{v} \rangle^2 \leq 1/(2t) \|\mathbf{v}\|_2^2$. Since this holds for all values of t, taking $t \to \infty$ tells us that $\langle \mathbf{u}, \mathbf{v} \rangle = 0$. This proves the Cauchy-Schwartz inequality for semi-inner products as well. □

B A Proof of Riesz Representation Theorem for Separable Hilbert Spaces

It is tempting to visualize what $\mathcal{F}_{\text{lin}}$ might look like since it is a collections of functionals over a vector space. In the exercises, we will show tha this collection actually forms a vector space in its own right. However, to visualize this vector space, we need to imagine what all kinds of linear functional might there exist over an arbitrary vector space V.

Now there is a class of linear functionals that we can immediately propose whenever the vector space we are working with is an inner product space. We will call this class $\mathcal{G}_{\text{lin}}$.

$$\mathcal{G}_{\text{lin}} = \{f_{\mathbf{v}} : \mathbf{u} \mapsto \langle \mathbf{u}, \mathbf{v} \rangle , \mathbf{v} \in V\}.$$

It is clear from the definition that every member of the class $\mathcal{G}_{\text{lin}}$ is indeed a bounded linear functional over V. Thus, it is clear that $\mathcal{G}_{\text{lin}} \subseteq \mathcal{F}_{\text{lin}}$. The Riesz representation theorem tells us for *Hilbert spaces*, the reverse is true as well.

Theorem 2.6. (*Riesz Representation Theorem*) Let $\mathcal{F}_{\text{lin}}$ and $\mathcal{G}_{\text{lin}}$ be be defined as above for a Hilbert space $\mathcal{H}$. Then we have $\mathcal{G}_{\text{lin}} = \mathcal{F}_{\text{lin}}$.

This result tells us that the method used to define $\mathcal{G}_{\text{lin}}$ is the only way bounded linear functionals can be defined. Proving this result in all its generality is beyond the scope of this course. However, to give a flavor of the result, we prove the result below for Hilbert spaces that have a countable orthonormal basis. For simplicity, we give the proof for finite basis, but the same may be extended to the case of a countably infinite basis as well. We note, however that the Riesz representation theorem is a much more elegant result that does not require a basis to be established over the Hilbert space.

Proof. To prove the result, we need to prove $\mathcal{G}_{\text{lin}} \supseteq \mathcal{F}_{\text{lin}}$. Let $V = \mathbb{R}^n$ be the n-dimensional Euclidean space with a finite orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ i.e. $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \mathbb{I}\{i = j\}$. Then every vector $\mathbf{u} \in V$ can be written as a unique linear combination of basis vectors of V as

$$\mathbf{u} = u_1 \cdot \mathbf{e}_1 + u_2 \cdot \mathbf{e}_2 + \ldots + u_n \cdot \mathbf{e}_n,$$

where $u_i = \langle \mathbf{u}, \mathbf{e}_i \rangle$. Now, given a linear functional $f \in \mathcal{F}_{\text{lin}}$, define the vector

$$\mathbf{v}_f = f(\mathbf{e}_1) \cdot \mathbf{e}_1 + f(\mathbf{e}_2) \cdot \mathbf{e}_2 + \ldots + f(\mathbf{e}_n) \cdot \mathbf{e}_n$$

Consider any vector $\mathbf{u} \in V$. We have

$$\begin{aligned} f(\mathbf{u}) &= f(u_1 \cdot \mathbf{e}_1 + u_2 \cdot \mathbf{e}_2 + \ldots + u_n \cdot \mathbf{e}_n) \\ &= f(u_1 \cdot \mathbf{e}_1) + f(u_2 \cdot \mathbf{e}_2) + \ldots + f(u_n \cdot \mathbf{e}_n) \\ &= u_1 \cdot f(\mathbf{e}_1) + u_2 \cdot f(\mathbf{e}_2) + \ldots + u_n \cdot f(\mathbf{e}_n) \\ &= \langle \mathbf{u}, \mathbf{e}_1 \rangle \cdot f(\mathbf{e}_1) + \langle \mathbf{u}, \mathbf{e}_2 \rangle \cdot f(\mathbf{e}_2) + \ldots + \langle \mathbf{u}, \mathbf{e}_n \rangle \cdot f(\mathbf{e}_n) \\ &= \langle \mathbf{u}, \mathbf{v}_f \rangle, \end{aligned}$$

which establishes the result. $\square$