Representing data in the form of vectors lies at the core of machine learning, data science, and scientific computing. These notes explain the basic theory of vector spaces over the real numbers. Differing from most introductory courses on linear algebra, we will adopt a “coordinate-free” viewpoint that treats vectors are an abstract data type supporting the operations of addition and scalar multiplication.

1 Algebraic definitions

Definition 1.1. A *vector space* (over the real numbers) is a non-empty set V of elements, called *vectors*, equipped with two operations, called *addition* and *scalar multiplication*.

- Addition is a binary operation of type $V \times V \to V$. In other words two vectors $\mathbf{x}$ and $\mathbf{y}$ can be added to yield another vector, $\mathbf{x} + \mathbf{y}$.
- Scalar multiplication is a binary operation of type $\mathbb{R} \times V \to V$. In other words we can scale a vector $\mathbf{x}$ by a real number a to obtain another vector, $a\mathbf{x}$.

These operations are required to satisfy the associative, commutative, distributive, and multiplicative identity laws.

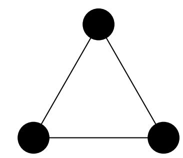
1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$.
2. $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$ and $(ab)\mathbf{x} = a(b\mathbf{x})$.
3. $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$ and $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$.
4. $1\mathbf{x} = \mathbf{x}$.

These laws imply the existence of a vector called the *zero vector*, which we denote by $\mathbf{0}$, that satisfies $\mathbf{0} + \mathbf{x} = \mathbf{x}$ and $0\mathbf{x} = \mathbf{0}$ for every $\mathbf{x} \in V$.

The most important and archetypical vector spaces are the spaces $\mathbb{R}^n$, defined for each positive integer n. Vectors in $\mathbb{R}^n$ are n-tuples of real numbers. Addition and scalar multiplication are defined component-wise. In these notes we will represent elements of $\mathbb{R}^n$ by column vectors, such as $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$.

The key distinction here is that we are not defining vectors to be n-tuples of real numbers, and then defining addition and scalar multiplication as operations on n-tuples. Rather, we are defining a vector space to be *any* structure possessing addition and scalar multiplication operations that satisfy the key properties in Definition 1.1, and then we are admitting the vector space $\mathbb{R}^n$ as an example of one such structure. This is similar to the distinction between an abstract data type, such as a list, and a data structure that implements that abstract data type, such as a doubly linked list. For the purpose of reasoning about vectors, everything we need to know about them is summarized in the abstract definition; for the purpose of calculating with them, we need to choose a specific way of representing the elements of a vector space, e.g., as n-tuples of real numbers.

Example 1.1. For any set S, there is a vector space $\mathbb{R}^S$ of functions from S to the real numbers, with addition and scalar multiplication defined pointwise: if $\mathbf{x}$ and $\mathbf{y}$ are two functions from S to $\mathbb{R}$, $a \in \mathbb{R}$ is a scalar, and s is any element of S, then $\mathbf{x} + \mathbf{y}$ is the function defined by $(\mathbf{x} + \mathbf{y})(s) = \mathbf{x}(s) + \mathbf{y}(s)$ and $a\mathbf{x}$ is the function defined by $(a\mathbf{x})(s) = a\mathbf{x}(s)$. For example, if G is the graph shown at right, then $\mathbb{R}^{V(G)}$ is the vector space of functions that label each vertex of G with a real number. It's evident that we can represent elements of $\mathbb{R}^{V(G)}$ as ordered 3-tuples of real numbers by choosing an ordering of the vertices of G. However, the choice of ordering is arbitrary, so there are at least six equally reasonable ways to model the elements of $\mathbb{R}^{V(G)}$ as elements of $\mathbb{R}^3$. We describe this state of affairs by saying that the vector spaces $\mathbb{R}^{V(G)}$ and $\mathbb{R}^3$ are *isomorphic* but not equal to one another.



Example 1.2. Continuing with the example above, let Z denote the subset of $\mathbb{R}^{V(G)}$ consisting of functions that sum to zero. In other words $\mathbf{x}$ belongs to Z if and only if it satisfies $\sum_{v \in V(G)} \mathbf{x}(v) = 0$. Then Z is also a vector space. An element of Z could be represented by an ordered triple of real numbers that sum to zero, such as the function values at the top, left, and right vertices respectively. Alternatively, we could represent an element of Z by an ordered pair of numbers, such as the function values at the left and right vertices only, since the value at the top vertex is uniquely determined by the other two. The vector space Z will become a running example in these notes.

1.1 Linear transformations and isomorphisms

Now that we've defined vector spaces, it's time to talk about functions between vector spaces. The most important type of function between vector spaces is called a linear transformation, and it preserves all of the algebraic structure of the space.

Definition 1.2. If V and W are vector spaces, a *linear transformation* from V to W is a function $T : V \to W$ that satisfies

$$T(a\mathbf{x} + b\mathbf{y}) = aT(\mathbf{x}) + bT(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in V$ and $a, b \in \mathbb{R}$.

A linear transformation is called an *isomorphism*, or equivalently *invertible*, if there is another linear transformation $T^{-1} : W \to V$ such that $T^{-1} \circ T$ and $T \circ T^{-1}$ are the identity functions of V and W, respectively. We then call T^{-1} the *inverse* of T. We say V and W are *isomorphic* if there is an isomorphism from V to W.

Example 1.3. When $m < n$, an important class of linear transformations from $\mathbb{R}^n$ to $\mathbb{R}^m$ are the *coordinate projections*: functions that modify an n-tuple to an m-tuple by extracting a specified subset of the coordinates. For example, the coordinate projection π_{13} from $\mathbb{R}^3$ to $\mathbb{R}^2$ deletes the middle coordinate of a 3-tuple while preserving the first and third coordinates, e.g. $\pi_{13}\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

For any vector spaces V and W, the set of linear transformations from V to W forms a vector space under pointwise addition and scalar multiplication. This vector space is denoted by $\hom(V, W)$. The isomorphisms from V to W don't form a vector space because, for example, when we multiply an isomorphism by the scalar 0 we obtain the function that maps every $\mathbf{x}$ in V to $\mathbf{0}$ in W, which is no longer an isomorphism.

1.2 Bases and dimension

It seems self-evident that the vector spaces $\mathbb{R}^2$ and $\mathbb{R}^3$ are not isomorphic, because one of them is two-dimensional while the other is three-dimensional. How do we actually define dimension of a vector space? How do we confirm that vector spaces of different dimensions are really not isomorphic to one another? To answer these questions, we must first introduce the very important notion of a *basis* for a vector space.

Definition 1.3. A *linear combination* of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is a finite sum of the form $a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$. It is non-trivial if at least one of the coefficients a_i is not equal to zero. A set S of vectors is *linearly independent* if the zero vector cannot be expressed as a non-trivial linear combination of elements of S. A *basis* of a vector space is a maximal linearly independent set.

Lemma 1.1. *If B is a basis of a vector space V, then every element of V can be represented as a linear combination of elements of B. This representation is unique, up to a reordering of the summands.*

Proof. If $v \in V$ cannot be represented as a linear combination of elements of B, then $B \cup \{v\}$ is linearly independent, contradicting the maximality of B. Hence, every element of V can be expressed as a linear combination of elements of B. To see why the representation is unique, consider any $\mathbf{x} \in V$ and consider two representations

$$\mathbf{x} = a_1\mathbf{b}_1 + \cdots + a_m\mathbf{b}_m = a_1'\mathbf{b}_1' + \cdots + a_n'\mathbf{b}_n'.$$

Subtracting these two representations of $\mathbf{x}$ from one another, we obtain a representation of $\mathbf{0}$ as a linear combination of elements of B. Since B is linearly independent, all the coefficients in this linear combination must be zero. Hence, the two representations of $\mathbf{x}$ are identical, up to a reordering of the terms of the sum. □

Corollary 1.2. *If V is a vector space with a finite basis B, then the linear transformation $T : \mathbb{R}^B \to V$ defined by $T(f) = \sum_{\mathbf{b} \in B} f(\mathbf{b})\mathbf{b}$ is an isomorphism.*

Proof. By Lemma 1.1, for every $\mathbf{x} \in V$ there is a unique representation of the form $\mathbf{x} = \sum_{\mathbf{b} \in B} a_{\mathbf{b}}\mathbf{b}$. Let $C(\mathbf{x})$ be the function in $\mathbb{R}^B$ defined by $C(\mathbf{x})(\mathbf{b}) = a_{\mathbf{b}}$. We leave it as an exercise for the reader to verify that C is a linear transformation and that $C \circ T$ and $T \circ C$ are the identity functions of their respective vector spaces. □

The image of $\mathbf{x} \in V$ under the isomorphism $C : V \to \mathbb{R}^B$ is a B-tuple of real numbers. The elements of this tuple are called the *components* of $\mathbf{x}$ in the basis B. **??** shows that every finite-dimensional vector space V is isomorphic to $\mathbb{R}^n$ for some value of n, and we will soon see this value is unique. However, there are many isomorphisms from V to $\mathbb{R}^n$, corresponding to all the different ways to choose an (ordered) basis of V. For this reason, the components of a vector are only well-defined in contexts where an ordered basis has been specified.

Definition 1.4. The *standard basis* of $\mathbb{R}^n$ is the basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$, where $\mathbf{e}_i$ denotes a vector whose i^{th} coordinate is 1 and all other coordinates are zero.

Lemma 1.3. *A set of vectors $B \subset V$ is a basis if and only if every element of V can be uniquely expressed as a linear combination of elements of B.*

Proof. The "only if" direction was proven in Lemma 1.1. If $B \subset V$ is a subset having the property that every element of V can be uniquely expressed as a linear combination of elements of B, then in particular the only representation of $\mathbf{0}$ as a linear combination of elements of B is the trivial representation; this verifies that B is linearly independent. Furthermore, for any $\mathbf{x} \notin B$, by our assumption on B we can find a representation $\mathbf{x} = a_1\mathbf{b}_1 + \cdots + a_m\mathbf{b}_m$. Then the equation $\mathbf{0} = a_1\mathbf{b}_1 + \cdots + a_m\mathbf{b}_m - \mathbf{x}$ shows that $B \cup \{vx\}$ is not linearly independent. Hence, B is a *maximal* linearly independent set, i.e. B is a basis, as claimed. □

We will be defining the dimension of a vector space to be the cardinality of any basis. However, in order to make such a definition we need to verify that all bases have the same cardinality. This is accomplished in the following pair of lemmas.

Lemma 1.4 (Exchange Lemma)**.** *If V is a vector space with basis B, then for any nonzero vector $\mathbf{x} \notin B$ we can obtain another basis from B by exchanging $\mathbf{x}$ for one of the vectors $\mathbf{y} \in B$. In other words, $B' = (B \setminus \{\mathbf{y}\}) \cup \{\mathbf{x}\}$.*

Proof. Using Lemma 1.1 and the fact that $\mathbf{x} \neq \mathbf{0}$, we know that $\mathbf{x}$ can be expressed as a non-trivial linear combination $\mathbf{x} = a_1\mathbf{b}_1 + \cdots + a_m\mathbf{b}_m$. Assume without loss of generality that $a_1 \neq 0$. Then

$$\mathbf{b}_1 = \mathbf{x} - \frac{a_2}{a_1}\mathbf{b}_2 - \cdots - \frac{a_m}{a_1}\mathbf{b}_m. \tag{1}$$

For any vector $\mathbf{z}$ that can be expressed as a linear combination of elements of B, we can substitute the right side of (1) in place of $\mathbf{b}_1$, to obtain a representation of $\mathbf{z}$ as a linear combination of elements of $B' = (B \setminus \{\mathbf{b}_1\}) \cup \{\mathbf{x}\}$. To see that this representation of $\mathbf{z}$ is unique, consider subtracting any two distinct representations of $\mathbf{z}$ as linear combinations of elements of B', to obtain a nontrivial representation of $\mathbf{0}$ as a linear combination of elements of B'. Let a_x denote the coefficient of $\mathbf{x}$ in this representation of $\mathbf{0}$. Our hypothesis that B is linearly independent means that $\mathbf{0}$ cannot be represented as a nontrivial linear combination of elements of B, so we know that $a_x \neq 0$. Now if we substitute the expression $a_1\mathbf{b}_1 + \cdots + a_m\mathbf{b}_m$ in place of $\mathbf{x}$, we obtain another representation of $\mathbf{0}$, this time as a linear combination of elements of B, in which the coefficient of $\mathbf{b}_1$ is a_1a_x. Since $a_1a_x \neq 0$, this contradicts the assumption that B is linearly independent. □

Theorem 1.5. *If V is a vector space with a finite basis, then all bases of V have the same number of elements.*

Proof. Let B and B' be two bases of V, with B finite. Denote the elements of B by $\{\mathbf{b}_1, \ldots, \mathbf{b}_d\}$. Now construct a sequence of bases by the following procedure. Start with $B_0 = B'$, and repeatedly perform the exchange procedure in the proof of Lemma 1.4, inserting elements of B one by one. This yields a sequence of bases $B_0, B_1, \ldots, B_d$, such that $B_0 = B'$, and for $i > 0$, $B_i = (B_{i-1} \cup \{\mathbf{b}_i\}) \setminus \mathbf{b}'_{i-1}$ for some $\mathbf{b}'_{i-1} \in B_{i-1}$. When choosing the vector $\mathbf{b}'_{i-1}$ to remove from B_{i-1} while inserting $\mathbf{b}_i$, let us *never remove a vector that belongs to B*. This is possible because in the proof of Lemma 1.4, the vector that was removed from the basis when inserting $\mathbf{x}$ was allowed to be any vector having a nonzero coefficient when $\mathbf{x}$ is represented using the basis B. We know that when $\mathbf{b}_i$ is represented using the basis B_{i-1}, at least one of the basis vectors with a nonzero coefficient does not belong to B; this is because B is linearly independent, so $\mathbf{b}_i$ cannot be represented as a non-trivial linear combination of elements of $B \setminus \{\mathbf{b}_i\}$.

By the time we reach B_d in this repeated-exchange process, we have inserted each element of B and have not removed any elements of B, so $B \subseteq B_d$. One basis cannot be a proper subset of another, since that would violate the maximality property of bases. Hence $B = B_d$. Since each two consecutive sets in the sequence of $B_0, \dots, B_d$ have the same cardinality, we conclude that $B' = B_0$ must have the same cardinality as B, as claimed. □

1.3 Inner products and the dual of a vector space

An important binary operation on $\mathbb{R}^n$ is the *dot product* operation, defined by $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$. In the setting of abstract vector spaces, the appropriate generalization of the dot product is a structure called a *positive definite inner product*, whose essential properties are defined as follows.

Definition 1.5. An *inner product* structure on a vector space is a function of type $V \times V \to \mathbb{R}$, with the inner product of vectors $\mathbf{x}, \mathbf{y} \in V$ being denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$. An inner product is required to satisfy the following properties.

1. Bilinearity:
$$\langle a\mathbf{x} + b\mathbf{y}, \mathbf{z} \rangle = a \langle \mathbf{x}, \mathbf{z} \rangle + b \langle \mathbf{y}, \mathbf{z} \rangle \quad \text{and} \quad \langle \mathbf{x}, a\mathbf{y} + b\mathbf{z} \rangle = a \langle \mathbf{x}, \mathbf{y} \rangle + b \langle \mathbf{x}, \mathbf{z} \rangle .$$
2. Symmetry:
$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle .$$

An inner product is called *non-degenerate* if for every $\mathbf{x} \neq \mathbf{0}$ there exists a $\mathbf{y}$ such that $\langle \mathbf{x}, \mathbf{y} \rangle \neq 0$. It is called *positive semidefinite* if $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for all $\mathbf{x}$, and *positive definite* if the inequality is strict for all $\mathbf{x} \neq \mathbf{0}$.

Note that a positive definite inner product is always non-degenerate: if $\mathbf{x} \neq \mathbf{0}$ then $\langle \mathbf{x}, \mathbf{x} \rangle \neq 0$. The dot product on $\mathbb{R}^n$ is positive definite because $\langle \mathbf{x}, \mathbf{x} \rangle = x_1^2 + \cdots + x_n^2$, which is always non-negative and equals zero only when $\mathbf{x} = \mathbf{0}$.

An example of a non-degenerate inner product that is *not* positive definite is the Lorentzian inner product on $\mathbb{R}^n$:
$$\langle \mathbf{x}, \mathbf{y} \rangle_L = -x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$
This inner product plays an important role in the physics of spacetime, where the first coordinate represents the time dimension and the remaining coordinates represent the spatial dimensions. According to Einstein's theory of special relativity, the linear transformations that one should apply to shift from one observer's system of spacetime coordinates to another's are precisely the linear transformations that preserve the Lorentzian inner product of vectors.

1.4 The dual of a vector space

Definition 1.6. The vector space $\hom(V, \mathbb{R})$ of real-valued linear functions on V is called the *dual* of V and is denoted by V^*.

Lemma 1.6. *Every finite-dimensional vector space is isomorphic to its own dual.*

Proof. Suppose V is a vector space and B is a finite basis for V. Recall from Corollary 1.2 that V is isomorphic to $\mathbb{R}^B$. The dual vector space V^* is also isomorphic to $\mathbb{R}^B$, via the isomorphism that maps each linear function $\mathbf{f} : V \to \mathbb{R}$ to the function $\mathbf{f}^B : B \to \mathbb{R}$ obtained by restricting the domain of $\mathbf{f}$ from V to B. To prove this is an isomorphism between V^* and B we need to prove it has an inverse. In other words, we need to show that for each function $\mathbf{f}^B : B \to \mathbb{R}$ there is a unique linear function $\mathbf{f} : V \to \mathbb{R}$ that restricts to $\mathbf{f}^B$. If $\mathbf{f}$ is any linear function that restricts to $\mathbf{f}^B$, then for any vector $\mathbf{x} = \sum_{\mathbf{b} \in B} x_{\mathbf{b}} \mathbf{b}$ the value $\mathbf{f}(\mathbf{x})$ must satisfy

$$\mathbf{f}(\mathbf{x}) = \sum_{\mathbf{b} \in B} x_{\mathbf{b}} \mathbf{f}^B(\mathbf{b}).$$

This shows there can be *at most one* linear function $\mathbf{f} : V \to \mathbb{R}$ that restricts to $\mathbf{f}^B$, since the value $\mathbf{f}(\mathbf{x})$ on any $\mathbf{x}$ is uniquely determined by the equation above. To verify that there is exactly one linear function that restricts to $\mathbf{f}^B$, we must check that the function $\mathbf{f}$ defined above is linear. Suppose $\mathbf{x}, \mathbf{y} \in V$ and $r, s \in \mathbb{R}$. If $\mathbf{x} = \sum_{\mathbf{b} \in B} x_{\mathbf{b}} \mathbf{b}$ and $\mathbf{y} = \sum_{\mathbf{b} \in B} y_{\mathbf{b}} \mathbf{b}$ then

$$r\mathbf{x} + s\mathbf{y} = \sum_{\mathbf{b} \in B} (r x_{\mathbf{b}} + s y_{\mathbf{b}}) \mathbf{b}$$

so

$$\mathbf{f}(r\mathbf{x} + s\mathbf{y}) = \sum_{\mathbf{b} \in B} (r x_{\mathbf{b}} + s y_{\mathbf{b}}) \mathbf{f}^B(\mathbf{b}) = r \sum_{\mathbf{b} \in B} x_{\mathbf{b}} \mathbf{f}^B(\mathbf{b}) + s \sum_{\mathbf{b} \in B} y_{\mathbf{b}} \mathbf{f}^B(\mathbf{b}) = r\mathbf{f}(\mathbf{x}) + s\mathbf{f}(\mathbf{y})$$

which confirms that $\mathbf{f}$ is linear. □

Example 1.4. Every real-valued linear function $\mathbf{f}$ on $\mathbb{R}^3$ can be represented (uniquely) by a sequence of three coefficients a_1, a_2, a_3 such that

$$\mathbf{f}\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = a_1 x_1 + a_2 x_2 + a_3 x_3.$$

The dual of $\mathbb{R}^3$ is isomorphic to $\mathbb{R}^3$, under the isomorphism that maps a linear function $\mathbf{f}$ to the coefficient vector $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$.

Generalizing the previous example, the dual of $\mathbb{R}^n$ is isomorphic to $\mathbb{R}^n$ via the isomorphism that maps a linear function to its coefficient vector. To facilitate distinguishing between $\mathbb{R}^n$ and its dual, we will represent elements of $(\mathbb{R}^n)^*$ as row vectors rather than column vectors. In other words, we will prefer to represent the linear function $\mathbf{f}$ in Example 1.4 using the row vector $\mathbf{f} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$ rather than the column vector $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$. This notation is convenient because it means that the application of the function $\mathbf{f}$ to the vector $\mathbf{x}$ can simply be written as $\mathbf{f}\mathbf{x}$, using the rules for multiplying a 1-by-n matrix by an n-by-1 matrix.

Generalizing these examples still further, a non-degenerate inner product structure on a finite-dimensional vector space always allows one to define an isomorphism between V and V^*. However, it's important to note that there are many isomorphisms between V and V^*, and there's no particular way to single out one of them without singling out a non-degenerate inner product structure.

Lemma 1.7. *If V is a finite dimensional vector space and $\langle \cdot, \cdot \rangle$ is a non-degenerate inner product, then there is an isomorphism $T : V \to V^*$ where $T(\mathbf{x})$ is defined to be the linear function $\mathbf{t}_\mathbf{x}$ specified by the formula $\mathbf{t}_\mathbf{x}(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$.*

Proof. The bilinearity property of inner products ensures that the function T defined in the lemma statement is a linear function from V to V^*. It is injective because if $\mathbf{x}, \mathbf{y} \in V$ satisfy $T(\mathbf{x}) = T(\mathbf{y})$, then for all $\mathbf{z} \in V$ we have $\langle \mathbf{x} - \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle - \langle \mathbf{y}, \mathbf{z} \rangle = \mathbf{t}_\mathbf{x}(\mathbf{z}) - \mathbf{t}_\mathbf{y}(\mathbf{z}) = 0$. As $\langle \cdot, \cdot \rangle$ is non-degenerate, this implies $\mathbf{x} - \mathbf{y} = 0$ hence $\mathbf{x} = \mathbf{y}$. From Lemma 1.6, we know that V and V^* have the same dimension. We leave it as an exercise to the reader to verify that an injective map between finite-dimensional vector spaces of the same dimension must be an isomorphism. □

Example 1.5. Suppose V is the subspace of $\mathbb{R}^3$ consisting of vectors whose coordinates sum to zero, with the positive definite inner product structure given by

$$\left\langle \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \\ z_z \end{bmatrix} \right\rangle = x_1 x_2 + y_1 y_2 + z_1 z_2.$$

One element of V^* is the linear function $\mathbf{f}$ that sums the first two coordinates of a vector, i.e. the function $\mathbf{f}\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = x + y$. If we are representing elements of V by three-tuples $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ then $\mathbf{f}$ can be represented by the row vector $[\,1\ 1\ 0\,]$. However, since $-z = x + y$ for every $\begin{bmatrix} x \\ y \\ z \end{bmatrix} \in V$, the function $\mathbf{f}$ is also expressed by the formula $\mathbf{f}\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = -z$ and can be represented by the vector $[\,0\ 0\ {-1}\,]$.

This example underscores the importance of distinguishing between a vector space and its dual. The vector space $\mathbb{R}^3$ is isomorphic to its dual, however when we pass to a subspace of $\mathbb{R}^3$, the dual of the subspace is *not* a subspace of $(\mathbb{R}^3)^*$. Instead, it is a *quotient* of $(\mathbb{R}^3)^*$, i.e. a vector space whose elements are *equivalence classes of vectors* in $(\mathbb{R}^3)^*$.

2 Convexity and norms

One the wonderful things about vector spaces is that, although they are defined by algebraic operations, we can also reason about them using geometric notions like convexity, distance, and volume. In this section we develop some basic facts about these three notions.

2.1 Convex sets and functions

A subset of a vector space is convex if it contains the line segment joining any two of its points. This informal definition is formalized as follows.

Definition 2.1. If $F = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ is a finite subset of a vector space V, an *affine combination* of points of F is a linear combination $a_1\mathbf{x}_1 + \cdots + a_m\mathbf{x}_m$ whose coefficients satisfy $a_1 + \cdots + a_m = 1$. A *convex combination* of points of F is an affine combination whose coefficients are non-negative. (Another name for a convex combination of vectors is a *weighted average*.) The *affine hull* and *convex hull* of F are the set of all affine combinations and all convex combinations of elements of F, respectively.

The affine hull of two points is the line passing through them, the affine hull of three non-collinear points is the plane passing through them, and so on. The convex hull of two points is the line segment joining them, the convex hull of three non-collinear points is the triangle joining them, and so on.

Definition 2.2. A subset K of a vector space is *convex* if every convex combination of points in K belongs to K. Equivalently, K is convex if, for every $\mathbf{x}, \mathbf{y} \in K$ and every $t \in [0, 1]$, the vector $(1 - t)\mathbf{x} + t\mathbf{y}$ also belongs to K.

A simple inductive proof verifies that the two formulations of convexity in Definition 2.2 are, indeed, equivalent. Clearly, the first definition implies the second because the expression $(1 - t)\mathbf{x} + t\mathbf{y}$ defines a convex combination of $\mathbf{x}$ and $\mathbf{y}$ when $0 \leq t \leq 1$. Conversely, suppose K satisfies the second definition. We assert that for every $m \geq 2$, every convex combination of m points of K belongs to K. The base case $m = 2$ is simply a restatement of the second definition. For the inductive step, if non-negative coefficients $a_1, a_2, \ldots, a_m$ sum up to 1, assume without loss of generality that $a_m > 0$, and let $t = 1 - a_1 = a_2 + a_3 + \cdots + a_m$. Since $t > 0$, we have

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_m\mathbf{x}_m = (1 - t)\mathbf{x}_1 + t\left(\tfrac{a_2}{t}\mathbf{x}_2 + \cdots + \tfrac{a_m}{t}\mathbf{x}_m\right).$$

By the induction hypothesis, the vector $\mathbf{x}' = \left(\frac{a_2}{t}\mathbf{x}_2 + \cdots + \frac{a_m}{t}\mathbf{x}_m\right)$ belongs to K. Hence, $(1-t)\mathbf{x}_1 + t\mathbf{x}'$ also belongs to K, as desired.

An important type of convex set is a halfspace, which is a set of the form

$$H = \{\mathbf{x} \mid \mathbf{f}(\mathbf{x}) \leq \theta\}, \tag{2}$$

for some nonzero $\mathbf{f} \in V^*$ and some $\theta \in \mathbb{R}$. Equivalently, due to Lemma 1.6, we can define a halfspace using a non-degenerate inner product as

$$H = \{\mathbf{x} \mid \langle \mathbf{w}, \mathbf{x} \rangle \leq b\}, \tag{3}$$

where $\mathbf{w}$ is a nonzero vector in V called the *normal vector* to H, and $\theta \in \mathbb{R}$. To verify that the set H defined using (2) is convex, observe that

$$\mathbf{f}(a_1\mathbf{x}_1 + \cdots + a_m\mathbf{x}_m) = a_1\mathbf{f}(\mathbf{x}_1) + \cdots + a_m\mathbf{f}(\mathbf{x}_m).$$

If $a_1, a_2, \ldots, a_m$ are the coefficients of a convex combination, then the right side of this equation is a weighted average of the values $\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_m)$. If each of those values is less than or equal to θ, then their weighted average must also be less than or equal to θ.

Convexity of a closed set[1] can be equivalently defined using halfspaces.

[1] A subset S of a finite-dimensional vector space is called *closed* if the limit point of every convergent sequence of vectors in S is also contained in S.

Lemma 2.1. *If V is a finite-dimensional vector space and K is a closed subset of V, then K is convex if and only if it is equal to the intersection of a set of halfspaces.*

The proof of the lemma is not quite self-contained. It uses some facts from topology that we state here without proof.

1. If V is a finite-dimensional vector space and $\langle \cdot, \cdot \rangle$ is a positive definite inner product, then for any $\mathbf{x}$ the function $q(\mathbf{y}) = \langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle$ is continuous.

2. If S is a non-empty, closed, bounded subset of a finite-dimensional vector space and f is a continuous function on S, then there exists a point $\mathbf{z} \in S$ such that $f(\mathbf{z}) = \inf\{f(\mathbf{y}) \mid \mathbf{y} \in S\}$.

Proof. From the definition of a convex set, it is clear that an intersection of convex sets is convex. Conversely, if K is convex then we must prove it is the intersection of a set of halfspaces. Specifically, let $\mathcal{H}(K)$ denote the set of halfspaces that contain K as a subset, and let $K' = \cap_{H \in \mathcal{H}(K)} H$. (If $\mathcal{H}(K) = \emptyset$ then interpret this intersection to be the entirety of V.) Then K' is the intersection of a set of halfspaces, and we shall show that $K' = K$. The containment $K \subseteq K'$ is immediate from the definition of K'. To show that $K' \subseteq K$, we prove the reverse containment $V \setminus K \subseteq V \setminus K'$. In other words, if $\mathbf{x} \in V \setminus K$, we must find a halfspace H that contains K but not $\mathbf{x}$. Let $\langle \cdot, \cdot \rangle$ be a positive-definite inner product on V, and consider the continuous function $q(\mathbf{y}) = \langle \mathbf{y} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle$. Let $q_0 = \inf\{q(\mathbf{y}) \mid \mathbf{y} \in K\}$ and observe that $q_0 > 0$. The set $K_0 = \{\mathbf{y} \in K \mid q(\mathbf{y}) \leq q_0 + 1\}$ is non-empty, closed, and bounded, so there exists $\mathbf{z} \in K_0$ with $q(\mathbf{z}) = q_0$.

Now consider the set

$$H = \{\mathbf{y} \mid \langle \mathbf{z} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \geq q_0\} = \{\mathbf{y} \mid \langle \mathbf{z} - \mathbf{x}, \mathbf{y} \rangle \geq q_0 + \langle \mathbf{z} - \mathbf{x}, \mathbf{x} \rangle\}.$$

This is a halfspace, and $\mathbf{x} \notin H$ because $\langle \mathbf{z} - \mathbf{x}, \mathbf{x} - \mathbf{x} \rangle = 0 < q_0$. To conclude the proof we will show that $K \subseteq H$. For any $\mathbf{y} \in K$ consider the function

$$f(t) = q(\mathbf{z} + t(\mathbf{y} - \mathbf{z})) = \langle \mathbf{z} - \mathbf{x} + t(\mathbf{y} - \mathbf{z}), \mathbf{z} - \mathbf{x} + t(\mathbf{y} - \mathbf{z}) \rangle = q(\mathbf{z}) + 2t \langle \mathbf{z} - \mathbf{x}, \mathbf{y} - \mathbf{z} \rangle + t^2 \langle \mathbf{y} - \mathbf{z}, \mathbf{y} - \mathbf{z} \rangle .$$

For $0 \leq t \leq 1$ the vector $\mathbf{z} + t(\mathbf{y} - \mathbf{z}) = (1 - t)\mathbf{z} + t\mathbf{y}$ belongs to K, and we know that the minimum value of q on K is attained at $\mathbf{z}$, so the quadratic function $f(t)$ on the interval $0 \leq t \leq 1$ attains its minimum value at $t = 0$. Therefore, $f'(0) \geq 0$, which implies $\langle \mathbf{z} - \mathbf{x}, \mathbf{y} - \mathbf{z} \rangle \geq 0$. Now, we find that

$$\langle \mathbf{z} - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle = \langle \mathbf{z} - \mathbf{x}, \mathbf{y} - \mathbf{z} \rangle + \langle \mathbf{z} - \mathbf{x}, \mathbf{z} - \mathbf{x} \rangle \geq 0 + q_0$$

hence $\mathbf{y}$ satisfies the defining inequality of the halfspace H. As $\mathbf{y}$ was an arbitrary element of K, we have proven $K \subseteq H$ as desired. ◻

In addition to convex sets, another important notion is that of a *convex function.*

Definition 2.3. If V is a vector space, $K \subseteq V$ is a convex set, and $h : K \to \mathbb{R}$ is a function, we say that h is convex if it satisfies

$$h((1 - t)\mathbf{x} + t\mathbf{y}) \leq (1 - t)h(\mathbf{x}) + th(\mathbf{y}) \qquad \forall \mathbf{x}, \mathbf{y} \in K,\ 0 \leq t \leq 1$$

Analogous to the two equivalent definitions of a convex set, it is equivalent to say that h is convex if and only if, for all finite sets $F = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subseteq K$ and convex combinations $\mathbf{x} = a_1\mathbf{x}_1 + \cdots + a_m\mathbf{x}_m$, the inequality

$$h(\mathbf{x}) \leq a_1 h(\mathbf{x}_1) + \cdots + a_m h(\mathbf{x}_m).$$

This inequality (along with its generalization to integrals rather than finite sums) goes by the name of *Jensen's convex function inequality.*

We proceed to state two more definitions related to convex functions and then a lemma providing two equivalent characterizations of convexity.

Definition 2.4. If V is a vector space, $K \subseteq V$, and $h : K \to \mathbb{R}$, then the *epigraph* of h is the set of all pairs $(\mathbf{x}, y) \in V \times \mathbb{R}$ such that $y \geq h(\mathbf{x})$. For any $\mathbf{x} \in K$, the *subdifferential* of h at $\mathbf{x}$ is defined to be the set

$$\partial h(\mathbf{x}) = \{\mathbf{f} \in V^* \mid \mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) \leq h(\mathbf{y}) - h(\mathbf{x}) \; \forall \mathbf{y} \in K\}.$$

One can visualize the epigraph of h as an infinitely tall multidimensional bowl-shaped region sitting above the graph of h in $V \times \mathbb{R}$. To visualize what it means for $\mathbf{f}$ to belong to the subdifferential of h, note that the graph of the function $L_{\mathbf{f},\mathbf{x}}(\mathbf{y}) = \mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}) + h(\mathbf{x})$ is a hyperplane in $V \times \mathbb{R}$ and it touches the graph of h at the point $(\mathbf{x}, h(\mathbf{x}))$. If the graph of $L_{\mathbf{f},\mathbf{x}}$ is a *supporting hyperplane* of the epigraph of h (i.e., a hyperplane that touches the epigraph of h at least once point and lies (weakly) below it everywhere) then $\mathbf{f}$ belongs to the subdifferential $\partial h(\mathbf{x})$.

If V has a non-degenerate inner product, this defines an isomorphism between V^* and V. The image of $\partial h(\mathbf{x})$ under this isomorphism is a set of vectors called the *subgradient* of h at $\mathbf{x}$.

To relate epigraphs and subgradients to convexity, we need to define one more notion: open subsets of a finite-dimensional vector space. Intuitively, a subset $U \subseteq V$ is *open* if every point of U is completely surrounded by other points of U. For example, in the open-dimensional vector space $\mathbb{R}$, an open interval $(a, b) = \{x \mid a < x < b\}$ is open whereas a closed interval $[a, b] = \{x \mid a \leq x \leq b\}$ is not, because the endpoints of a closed interval are not surrounded on both sides by other points of the interval.

Definition 2.5. If V is a finite-dimensional vector space, a subset $U \subseteq V$ is called an *open set* if it satisfies the following property: for all $\mathbf{x}, \mathbf{y} \in U$ there exists some $\delta > 0$ such that for every ε with $|\varepsilon| < \delta$, the vector $\mathbf{x} + \varepsilon\mathbf{y}$ belongs to U.

Lemma 2.2. *For a convex open subset K of a finite-dimensional vector space V, the following properties of a function $h : K \to \mathbb{R}$ are equivalent.*

1. *h is convex.*
2. *The epigraph of h is a convex subset of $V \times \mathbb{R}$.*
3. *The subdifferential of h is nonempty at every point of K.*

Proof. We will prove the cycle of implications (3) ⇒ (1) ⇒ (2) ⇒ (3), which suffices to prove the equivalence of the three condiions.

(3) ⇒ (1): If the subdifferential of h is nonempty at every point of K, then consider any two points $\mathbf{x}, \mathbf{x}'$ and their convex combination $\mathbf{x}'' = (1-t)\mathbf{x} + t\mathbf{x}'$. The subdifferential $\partial h(\mathbf{x}'')$ is non-empty, so it contains some $\mathbf{f} \in V^*$ that satisfies $\mathbf{f}(\mathbf{y}) - \mathbf{f}(\mathbf{x}'') \le h(\mathbf{y}) - h(\mathbf{x}'')$ for all $\mathbf{y} \in K$. In particular, we have the two inequalities

$$\begin{aligned} \mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}'') &\le h(\mathbf{x}) - h(\mathbf{x}'') \\ \mathbf{f}(\mathbf{x}') - \mathbf{f}(\mathbf{x}'') &\le h(\mathbf{x}') - h(\mathbf{x}''). \end{aligned}$$

Multiplying the first by $1-t$ and the second by t we obtain

$$(1-t)\mathbf{f}(\mathbf{x}) + t\mathbf{f}(\mathbf{x}') - \mathbf{f}(\mathbf{x}'') \le (1-t)h(\mathbf{x}) + th(\mathbf{x}') - h(\mathbf{x}'').$$

The left side is zero, because $\mathbf{f}$ is a linear function that $\mathbf{x}'' = (1-t)\mathbf{x}+t\mathbf{x}'$. Hence, $(1-t)h(\mathbf{x})+th(\mathbf{x}') \ge h(\mathbf{x}'') = h((1-t)\mathbf{x} + t\mathbf{x}'')$ which confirms that h is convex.

(1) ⇒ (2): Suppose h is convex. Let $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ denote two points in the epigraph of h. Then $y \ge h(\mathbf{x})$ and $y' \ge h(\mathbf{x}')$ so

$$(1-t)y + ty' \ge (1-t)h(\mathbf{x}) + th(\mathbf{x}') \ge h((1-t)\mathbf{x} + t\mathbf{x}')$$

which shows that $(1-t)(\mathbf{x}, y) + t(\mathbf{x}', y')$ belongs to the epigraph of h and thus confirms that the epigraph is convex.

(2) ⇒ (3): If the epigraph of h is convex and $\mathbf{x}$ is a point of K, then for every $n > 0$ the point $(\mathbf{x}, h(\mathbf{x}) - 1/n)$ does not belong to the epigraph of h. The closure of the epigraph of h (i.e., the set consisting of the epigraph along with every point in $V \times \mathbb{R}$ that is the limit of a sequence of points in the epigraph) is a closed, convex subset of $V \times \mathbb{R}$. By Lemma 2.1 it follows that there is a halfspace H_n that contains the epigraph of h but doesn't contain $(\mathbf{x}, h(\mathbf{x}) - 1/n)$. The set of points $(\mathbf{x}', y') \in H_n$ is defined by an inequality of the form $\mathbf{f}_n(\mathbf{x}') - a_n y' \le \theta_n$, where $\mathbf{f}_n \in V^*$ and $a_n, \theta_n \in \mathbb{R}$.

Choose an isomorphism between $V^* \times \mathbb{R}$ and $\mathbb{R}^{d+1}$, where d is the dimension of V, and let S be the image of the unit sphere in $\mathbb{R}^{d+1}$ under this isomorphism. By rescaling $(\mathbf{f}_n, a_n)$ if necessary, we can assume that $(\mathbf{f}_n, a_n) \in S$ for each n. Since S is a closed and bounded subset of $V^* \times \mathbb{R}$, and $V^* \times \mathbb{R}$ is finite dimensional, the sequence $(\mathbf{f}_n, a_n)_{n=1}^{\infty}$ has an infinite subsequence that converges to a limit point $(\mathbf{f}, a) \in S$. If we look at the values θ_n as n ranges over the same subsequence, we claim that they converge to the number $\theta = \mathbf{x}(\mathbf{x}) - ah(\mathbf{x})$. To prove this, note that $(\mathbf{x}, h(\mathbf{x})) \in H_n$ but $(\mathbf{x}, h(\mathbf{x}) - 1/n) \notin H_n$, which means

$$\mathbf{f}_n(\mathbf{x}) - a_n h(\mathbf{x}) \le \theta_n < \mathbf{f}_n(\mathbf{x}) - a_n h(\mathbf{x}) + a_n/n. \tag{4}$$

Passing to a subsequence on which $(\mathbf{f}_n, a_n)$ converges to $(\mathbf{f}, a)$, the left and right sides both converge to $\mathbf{f}(\mathbf{x}) - ah(\mathbf{x})$, so θ_n must converge to the same number.

Next we claim that the halfspace H consisting of all pairs $(\mathbf{x}', y')$ satisfying the inequality $\mathbf{f}(\mathbf{x}') - ay' \le \theta$ contains the epigraph of h. To see that this is so, assume $y' \ge h(\mathbf{x}')$ and note that for each n we know that $(\mathbf{x}', y')$ belongs to H_n hence it satisfies $\mathbf{f}_n(\mathbf{x}') - a_n y' \le \theta_n$. Passing to a subsequence on which $(\mathbf{f}_n, a_n) \to (\mathbf{f}, a)$ and then taking the $\liminf$ of both sides, we find that $\mathbf{f}(\mathbf{x}') - ay' \le \theta$, as claimed.

For any $\mathbf{x}'$ in K, the point $(\mathbf{x}', h(\mathbf{x}'))$ belongs to the epigraph of H, hence it satisfies

$$\mathbf{f}(\mathbf{x}') - ah(\mathbf{x}') \leq \theta = \mathbf{f}(\mathbf{x}) - ah(\mathbf{x}).$$

Rearranging this equation we find that

$$\tfrac{1}{a}\mathbf{f}(\mathbf{x}') - \tfrac{1}{a}\mathbf{f}(\mathbf{x}) \leq h(\mathbf{x}') - h(\mathbf{x})$$

for all $\mathbf{x}' \in K$, which confirms that $\frac{1}{a}\mathbf{f}$ belongs to ∂h, so ∂h is nonempty. □

2.2 Norms

A *norm* on a vector space provides a way to measure the length of a vector, and hence the distance between two vectors.

Definition 2.6. If V is a vector space, a *norm* on V is a function $\|\cdot\|$ from V to $\mathbb{R}$ satisfying:

1. Non-negativity: $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in V$, with equality if and only if $\mathbf{x} = 0$.
2. Linear homogeneity: $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$ for all $a \in \mathbb{R}$ and $\mathbf{x} \in V$.
3. Subadditivity: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in V$.

Common examples of norms on $\mathbb{R}^n$ are the L_p norms, defined for $1 \leq p < \infty$ by

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

and for $p = \infty$ by

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

It is easy to check that these norms satisfy non-negativity and linear homogeneity; the proof of subadditivity is omitted from these notes but can be found in many textbooks.

Lemma 2.3. *For $\mathbf{x} \in \mathbb{R}^n$, the p-norm $\|\mathbf{x}\|_p$ is a non-increasing function of p.*

Proof. For $\mathbf{x} = \mathbf{0}$ the assertion is trivial, since $\|\mathbf{x}\|_p = 0$ for all p. Otherwise, consider any $\mathbf{x} \neq \mathbf{0}$ and any p, q such that $1 \leq p < q$. We wish to show that $\|\mathbf{x}\|_p \geq \|\mathbf{x}\|_q$. By rescaling $\mathbf{x}$ if necessary, we may assume $\|\mathbf{x}\|_q = 1$. (The rescaling doesn't affect the validity of the inequality, since the linear homogeneity property ensures both sides are scaled by the same amount.) This implies that $|x_i| \leq 1$ for all i, either because $q = \infty$ or because $q < \infty$, $\sum_{i=1}^{q} |x_i|^q \leq 1$, and every term in the sum is non-negative. Since $|x_i| \leq 1$ and $p < q$, we have $|x_i|^p \geq |x_i|^q$. Summing these inequalities,

$$\|\mathbf{x}\|_p^p = \sum_{i=1}^{n} |x_i|^p \geq \sum_{i=1}^{n} |x_i|^q = 1.$$

Taking the p^{th} root of both sides, $\|\mathbf{x}\|_p \geq 1 = \|\mathbf{x}\|_q$. □

When $\mathbf{x}$ is a vector with just one nonzero coordinate x_i, the p-norm $\|\mathbf{x}\|_p$ is equal to $|x_i|$ for every p. When $\mathbf{x}$ has more than one nonzero coordinate, $\|\mathbf{x}\|_p$ is a strictly decreasing function of p: it is largest when $p = 1$ and smallest when $p = \infty$. More generally, having large 1-norm can often be interpreted as a sign of *density* (i.e., having many nonzero coordinates) while having small 1-norm is often interpreted as a sign of *sparsity.* This intuition will be put to use later in the course.

Definition 2.7. If V is a vector space and $\|\cdot\|$ is a norm, the *unit ball* of $\|\cdot\|$ is the set of all vectors in V whose norm is less than or equal to 1.

Lemma 2.4. *If V is a vector space and $\|\cdot\|$ is a norm, the unit ball of $\|\cdot\|$ is a closed, bounded, convex set that is* centrally symmetric, *meaning that for every vector $\mathbf{x}$ in the unit ball, $-\mathbf{x}$ along belongs to the unit ball. Conversely, for any closed, bounded, centrally symmetric convex set B, there exists a norm whose unit ball is B.*

The following important inequality is usually called the Cauchy-Schwartz inequality.

Lemma 2.5. *If $\langle \cdot, \cdot \rangle$ is a positive definite inner product on a vector space, then for any two vectors $\mathbf{x}, \mathbf{y}$ we have*

$$\langle \mathbf{x}, \mathbf{y} \rangle \le \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} \cdot \langle \mathbf{y}, \mathbf{y} \rangle^{1/2} ,$$

with equality if and only if $\mathbf{x}$ is a scalar multiple of $\mathbf{y}$ or vice-versa.

Proof. If $\mathbf{x}$ or $\mathbf{y}$ is equal to $\mathbf{0}$ then both sides of the inequality are zero, so the lemma holds. Otherwise, note that replacing $\mathbf{x}$ and $\mathbf{y}$ with $a\mathbf{x}$ and $b\mathbf{y}$, respectively, multiplies both sides of the inequality by ab. Hence, we may prove the lemma in the special case when $\langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle = 1$; the general case will then follow by scaling $\mathbf{x}$ and $\mathbf{y}$ suitably.

When $\langle \mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{y}, \mathbf{y} \rangle = 1$, we have

$$0 \le \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle - 2 \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle = 2 - 2 \langle \mathbf{x}, \mathbf{y} \rangle .$$

Furthermore, the inequality is strict when $\mathbf{x} - \mathbf{y} \neq \mathbf{0}$. Hence, we conclude that $\langle \mathbf{x}, \mathbf{y} \rangle \le 1 = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2} \cdot \langle \mathbf{y}, \mathbf{y} \rangle^{1/2}$ and that the inequality is strict unless $\mathbf{x} = \mathbf{y}$. □

An easy application of the Cauchy-Schwartz inequality shows that any positive definite inner product can be used to define a norm on a vector space.

Lemma 2.6. *If V is a vector space with a positive definite inner product $\langle \cdot, \cdot \rangle$, then the function defined by*

$$\|\mathbf{x}\| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$$

is a norm.

Proof. Non-negativity follows from positive definiteness of the inner product, and linear homogeneity follows from bilinearity. To prove subadditivity, observe that for any $\mathbf{x}, \mathbf{y}$,

$$\|\mathbf{x} + \mathbf{y}\|^2 = \langle \mathbf{x} + \mathbf{y}, \mathbf{x} + \mathbf{y} \rangle = \|\mathbf{x}\|^2 + 2 \langle \mathbf{x}, \mathbf{y} \rangle + \|\mathbf{y}\|^2.$$
$$(\|\mathbf{x}\| + \|vy\|)^2 = \|\mathbf{x}\|^2 + 2\|\mathbf{x}\| \, \|\mathbf{y}\| + \|\mathbf{y}\|^2.$$

The Cauchy-Schwartz inequality implies that the right side of the first equation is less than or equal to the right side of the second equation. □

For the standard inner product on $\mathbb{R}^n$ the norm defined in Lemma 2.6 coincides with the L_2 norm. For other positive definite inner products on $\mathbb{R}^n$, it constitutes a different norm whose unit ball is an ellipsoidal (egg-shaped) region.

2.3 Differentials and gradients

The gradient of a function on $\mathbb{R}^n$ is usually defined using partial derivatives. In this section we will see that a differentiable function on a vector space V always has a well-defined "differential" at every point, which is an element of the dual space V^*. However, to define the gradient requires choosing an isomorphism between V and V^*; hence, the gradient of a multivariate function depends on the choice of inner product structure for the vector space on which the function is defined.

Definition 2.8. If $(V, \|\cdot\|)$ is a normed vector space, a function $g : V \to \mathbb{R}$ is said to *vanish to first order at* $\mathbf{0}$ if $\frac{g(\mathbf{x})}{\|\mathbf{x}\|} \to 0$ as $\|\mathbf{x}\| \to 0$, uniformly in $\mathbf{x}$. More precisely, g vanishes to first order at $\mathbf{0}$ if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that $\frac{g(\mathbf{x})}{\|\mathbf{x}\|} < \varepsilon$ whenever $\|\mathbf{x}\| < \delta$.

Definition 2.9. If $(V, \|\cdot\|)$ is a normed vector space, $S \subseteq V$, and $f : V \to \mathbb{R}$, we say that f is *differentiable* at a point $\mathbf{x} \in S$ if there exists a linear function $\mathbf{df_x} \in V^*$, called the *differential* of f at $\mathbf{x}$, such that

$$\forall \mathbf{y}\ f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + \mathbf{df_x}(\mathbf{y}) + g(\mathbf{y}),$$

where the remainder $g(\mathbf{y})$ vanishes to first order at $\mathbf{0}$. If f is differentiable at every point of S, we simply say that f is differentiable.

The following lemma explains the relationship between differentials and subdifferentials of convex functions.

Lemma 2.7. *If f is a convex function and f is differentiable at $\mathbf{x}$, then the subdifferential $\partial f(\mathbf{x})$ at the point $\mathbf{x}$ is the one-element set $\{\mathbf{df_x}\}$.*

Proof. Let $g(\mathbf{y}) = f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x}) - \mathbf{df_x}(\mathbf{y})$. From the Definition 2.9 we know that g vanishes to first order at $\mathbf{0}$. On the other hand, g is convex because it is a convex function, minus a constant, minus a linear function. To complete the proof of the lemma it suffices to prove that the subdifferential of g at $\mathbf{0}$ is a singleton set consisting of $\mathbf{0} \in V^*$, i.e. the constant function that maps every vector in V to 0. From Definition 2.3 we know that the subdifferential $\partial g(\mathbf{x})$ is a nonempty set. To prove it equals $\{\mathbf{0}\}$, let $\mathbf{h}$ be any nonzero element of V^* and we will show $\mathbf{h} \notin \partial g(\mathbf{0})$. Suppose $\mathbf{y}$ is a vector such that $\mathbf{h}(\mathbf{y}) \neq 0$. Replacing $\mathbf{y}$ with $-\mathbf{y}$ if necessary, we can assume $\mathbf{h}(\mathbf{y}) > 0$. Now, since $\mathbf{h} \in \partial g(\mathbf{0})$, we have $g(\mathbf{z}) \geq g(\mathbf{0}) + \mathbf{h}(\mathbf{z} - \mathbf{0}) = \mathbf{h}(\mathbf{z})$ for all vectors $\mathbf{z}$. In particular, letting $\mathbf{z} = t\mathbf{y}$ for $t \in \mathbb{R}$, we find that $g(\mathbf{z}) = t\mathbf{h}(\mathbf{y})$ and

$$\lim_{t\to 0} \frac{g(\mathbf{z})}{\|\mathbf{z}\|} = \lim_{t\to 0} \frac{t\,h(\mathbf{y})}{t\,\|\mathbf{y}\|} = \frac{h(\mathbf{y})}{\|\mathbf{y}\|} > 0.$$

This contradicts the fact that g vanishes to first order at $\mathbf{0}$. □

Closely related to the differential of a function is its gradient, which encodes information about the derivative of f in the form of a vector in V rather than V^*.

Definition 2.10. If V is a vector space, $\langle \cdot, \cdot \rangle$ is a non-degenerate inner product, and $f : V \to \mathbb{R}$ is a function differentiable at $\mathbf{x}$, the *gradient* of f at $\mathbf{x}$, denoted by $\nabla \mathbf{f}_\mathbf{x}$, is the image of the differential $\mathbf{df}_\mathbf{x}$ under the isomorphism $V^* \to V$ induced by the inner product.

When $V = \mathbb{R}^n$ with the standard inner product structure, these definitions accord with the usual definitions given using partial derivatives. The differential of f is the row vector

$$\mathbf{df}_\mathbf{x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}$$

and the gradient $\nabla \mathbf{f}_\mathbf{x}$ is the column vector obtained by transposing this row vector.

Example 2.1. This example illustrates the difference between the gradient with respect to the standard inner product and the gradient with respect to a non-standard inner product. Let $V = \mathbb{R}^2$ and consider the function $f : V \to \mathbb{R}$ defined by $f(x_1, x_2) = 4x_1^2 + x_2^2$.

To calculate the differential of f at $\mathbf{x} = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$, we expand $f(\mathbf{x} + \mathbf{y})$ in powers of y_1 and y_2:

$$f(\mathbf{x}+\mathbf{y}) = 4(x_1+y_1)^2+(x_2+y_2)^2 = (4x_1^2+x_2^2) + (8x_1y_1+2x_2y_2) + (4y_1^2+y_2^2) = f(\mathbf{x})+(8x_1y_1+2x_2y_2)+g(\mathbf{y}),$$

where the function $g(\mathbf{y}) = 4y_1^2 + y_2^2$ vanishes to first order at $\mathbf{0}$. This indicates that

$$\mathbf{df}_\mathbf{x}(\mathbf{y}) = 8x_1y_1 + 2x_2y_2.$$

The right side of the equation is a linear function of $\mathbf{y} \in \mathbb{R}^2$. In other words, the differential of f an element of $(\mathbb{R}^2)^*$, as expected.

The gradient of f with respect to the standard inner product is obtained by stacking the two partial derivatives of f into a vector.

$$\nabla \mathbf{f}_\mathbf{x} = \begin{bmatrix} 8x_1 \\ 2x_2 \end{bmatrix}.$$

What about the gradient of f with respect to the non-standard inner product defined by

$$\langle \mathbf{x}, \mathbf{y} \rangle = 2x_1y_1 + x_2y_2.$$

The gradient $\nabla \mathbf{f}_\mathbf{x}$ is defined to be the image of $\mathbf{df}_\mathbf{x}$ under the isomorphism $(\mathbb{R}^2)^* \to \mathbb{R}^2$ induced by the inner product. In other words, $\nabla \mathbf{f}_\mathbf{x}$ is the unique vector $\mathbf{z} = \left[\begin{smallmatrix} z_1 \\ z_2 \end{smallmatrix}\right]$ that satisfies

$$\forall \mathbf{y} = \left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right] \quad \langle \mathbf{z}, \mathbf{y} \rangle = 8x_1y_1 + 2x_2y_2.$$

Recall that the inner product $\langle \mathbf{z}, \mathbf{y} \rangle$ is defined to be $2z_1y_1 + z_2y_2$. So, for all $\mathbf{y} \in \mathbb{R}^2$, we require the equation

$$8x_1y_1 + 2x_2y_2 = 2z_1y_1 + z_2y_2$$

to hold. Equating the coefficients of y_1 and y_2, we may conclude that $z_1 = 4x_1$ and $z_2 = 2x_2$. Hence,

$$\nabla \mathbf{f}_\mathbf{x} = \begin{bmatrix} 4x_1 \\ 2x_2 \end{bmatrix}.$$

2.4 Gradient descent

Minimizing a real-valued function on a vector space is one of the most important optimization problems in Computer Science. Among other uses, it underlies the training of machine learning models: in that application, each vector in the vector space represents a different parameter setting for the model, and the function to be minimized is called a “loss function” and is interpreted as a measure of how poorly the model with those parameters fits the training data.

The most popular family of algorithms for minimizing real-valued functions on vector spaces is based on a principle called *gradient descent*. These are iterative algorithms that take a sequence of small steps, each in a direction that locally improves the function value. In this section we introduce the gradient descent algorithm and analyze its performance when minimizing a convex function. Many of the most important contemporary applications of gradient descent involve non-convex functions, but the performance guarantees for gradient descent are much weaker when the function being optimized is non-convex.

The most elementary gradient descent algorithm has a “step size” parameter, η. The algorithm is as follows.

Algorithm 1 Gradient descent with fixed step size

Parameters: Starting point $\mathbf{x}_0 \in \mathbb{R}^n$, step size $\eta > 0$, number of iterations $T \in \mathbb{N}$.

for $t = 0, \ldots, T-1$ **do**
    $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}_t}$
end for
Output $\hat{\mathbf{x}} = \arg\min\{f(\mathbf{x}_0), \ldots, f(\mathbf{x}_T)\}$.

We will analyze the behavior of gradient descent under the following assumptions.

1. V has a positive definite inner product, $\langle \cdot, \cdot \rangle$. Gradients and norms of vectors are defined with respect to this inner product.

2. f is convex.

3. For some $L < \infty$ called the *Lipschitz constant of f*, the following inequality is satisfied by all $\mathbf{x}, \mathbf{y} \in V$.
$$|f(\mathbf{x}) - f(\mathbf{y})| \leq L \cdot \|\mathbf{x} - \mathbf{y}\|.$$

Let $\mathbf{x}^*$ denote a point in V at which f is minimized. The analysis of the algorithm will show that if $\|\mathbf{x}^* - \mathbf{x}_0\| \leq D$ then gradient descent (Algorithm 1) with $\eta = \varepsilon / L^2$ finds a point $\hat{\mathbf{x}}$ where $f(\hat{\mathbf{x}}) \leq f(\mathbf{x}^*) + \varepsilon$ after $T = L^2 D^2 / \varepsilon^2$ iterations. The key parameter in the analysis is the squared distance $\|\mathbf{x}_t - \mathbf{x}^*\|^2$. The following lemma does most of the work, by showing that this parameter must decrease if $f(\mathbf{x}_t)$ is sufficiently far from $f(\mathbf{x}^*)$.

Lemma 2.8. $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta(f(\mathbf{x}_t) - f(\mathbf{x}^*)) + \eta^2 L^2$.

Proof. Letting $\mathbf{x} = \mathbf{x}_t$ we have

$$\begin{aligned} \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x} - \mathbf{x}^* - \eta \boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}}\|^2 \\ &= \|\mathbf{x} - \mathbf{x}^*\|^2 - 2\eta \langle \boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}}, \mathbf{x} - \mathbf{x}^* \rangle + \eta^2 \|\boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}}\|^2 \\ &= \|\mathbf{x} - \mathbf{x}^*\|^2 + 2\eta \mathbf{df}_{\mathbf{x}}(\mathbf{x}^* - \mathbf{x}) + \eta^2 \|\boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}}\|^2 \\ &\le \|\mathbf{x} - \mathbf{x}^*\|^2 + 2\eta (f(x^*) - f(x)) + \eta^2 \|\boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}}\|^2. \end{aligned}$$

The proof concludes by observing that the L-Lipschitz property of f implies $\|\boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}}\| \le L$. □

Now, to complete the analysis of gradient descent, let $\Phi(t) = \|\mathbf{x}^t - \mathbf{x}^*\|^2$; we will refer to Φ as the "potential function" and to $\Phi(t)$ as the "potential at time t". When $\eta = \varepsilon/L^2$, the lemma implies that for every t such that $f(\mathbf{x}_t) > f(\mathbf{x}^*) + \varepsilon$, the decrease in potential at time t is bounded below by

$$\Phi(t) - \Phi(t+1) > 2\eta\varepsilon - \eta^2 L^2 = \varepsilon^2/L^2. \tag{5}$$

Since $\Phi(0) \le D$ and $\Phi(t) \ge 0$ for all t, the equation (5) cannot be satisfied for all $0 \le t \le L^2 D^2/\varepsilon^2$. Hence, if we run gradient descent for $T = L^2 D^2/\varepsilon^2$ iterations, at least one of the iterates $\mathbf{x}_t$ satisfies $f(\mathbf{x}_t) \le f(\mathbf{x}^*) + \varepsilon$, and hence the algorithm will set $\hat{\mathbf{x}}$ to be a point that satisfies such that $f(\hat{\mathbf{x}}) \le f(\mathbf{x}^*) + \varepsilon$.

A few observations about this analysis of gradient descent are in order.

1. The upper bound on the number of iterations *does not depend on the dimension of the vector space*. The bound is $L^2 D^2/\varepsilon^2$, which depends on the Lipschitz constant of the function (namely L) and on the distance of the starting point $\mathbf{x}_0$ from the optimal point $\mathbf{x}^*$ (namely D), but the number of iterations required to find an ε-optimal point does not tend to infinity as the dimension increases, provided those other parameters do not increase with dimension. This partially explains why gradient descent is such a useful algorithm for contemporary optimization problems with billions of parameters, such as training very large neural networks. To be honest, though, in those applications it is quite unlikely that the initial distance from optimality, D, would remain constant as the number of parameters tends to infinity.

2. The number of iterations depends quadratically on $1/\varepsilon$, which is quite bad. Later in the course we will see a variant of gradient descent that needs only $O(\log(1/\varepsilon))$ iterations, when the gradient $\boldsymbol{\nabla} \mathbf{f}_{\mathbf{x}}$ is neither too rapidly nor too slowly varying as $\mathbf{x}$ varies.

3. As noted in Section 2.3, the gradient (unlike the differential) is only well-defined in the context of an inner product structure on V. Under a different choice of inner product, the gradient of a function would be calculated in a different way, which would cause gradient descent to behave differently. This can be seen by plotting the iterations of gradient descent when minimizing a function such as $f(x) = 4x^2 + y^2$, whose level sets are ellipses. The gradient vectors with respect to the standard inner product are perpendicular to the level sets. The negative gradient (i.e., the direction of the steps taken by the gradient descent algorithm) is directed toward a point on the major axis of the ellipse but not toward its center. Hence, gradient descent with respect to the standard inner product will tend to zig-zag back and forth across the major axis as it makes it way toward the global minimum of

f, repeatedly overshooting in the x direction and then correcting its course, while making steady progress in the y direction. If, instead of the standard inner product, one takes the gradient of f with respect to the non-standard inner product

$$\langle x, y \rangle = 2x_1y_1 + x_2y_2,$$

then the gradient descent algorithm makes steady progress in both the x and y directions. Thus, while gradient descent using the standard inner product is adequately efficient, if one knows something about the geometry of the function being optimized then choosing an inner product adapted to the geometry of the problem can make gradient descent even more efficient.

3 Geometry in high dimensions

When visualizing high-dimensional vector spaces, it is important to keep in mind some stark quantitative differences between low-dimensional and high-dimensional geometry. In high dimensions, when we circumscribe a cube around a sphere, the cube's volume exceeds that of the sphere by a greater-than-exponential factor. (In other words, as the dimension increases, the volume ratio of the two shapes grows faster than any exponential function of the dimension.) Almost all of the volume of a high-dimensional ball is located in a thin shell near the surface. In addition, almost all of the ball's volume is located near the equator. Finally, if we sample m vectors at random from a d-dimensional ball and m is subexponential in d, then with high probability all of the vectors are nearly orthogonal to one another.

3.1 Preliminaries

We will derive all of the geometric facts cited above using a few basic facts from geometry and analysis.

In the vector space $\mathbb{R}^d$ there is a function denoted by $\text{Vol}_d(\cdot)$ that assigns to certain subsets $S \subseteq \mathbb{R}^d$ a non-negative (possibly infinite) number $\text{Vol}_d(S)$ called the d-dimensional volume of S. The sets for which $\text{Vol}_d(S)$ is defined are called *measurable sets* and we will not give a definition here, but we will note that any (topologically) closed subset of $\mathbb{R}^d$ is measurable, and the collection of measurable subsets is closed under complementation and under taking unions or intersections of countably many sets. The d-dimensional volume of a set S contained in a d-dimensional hyperplane in $\mathbb{R}^n$ (i.e., a set obtained from a d-dimensional linear subspace by translation) because

Furthermore, the d-dimensional volume satisfies the following properties.

1. The d-dimensional volume of a set is invariant under translations and rotations.
2. When we scale a set by a scale factor $\lambda > 0$, its d-dimensional volume is scaled by λ^d. In other words, if we define
$$\lambda \cdot S = \{\lambda \cdot \mathbf{x} \mid \mathbf{x} \in S\}$$
and if S is measurable, then $\lambda \cdot S$ is measurable and $\text{Vol}_d(\lambda \cdot S) = \lambda^d \cdot \text{Vol}_d(S)$.

3. If A and B are disjoint measurable sets, then $\mathrm{Vol}_d(A \cup B) = \mathrm{Vol}_d(A) + \mathrm{Vol}_d(B)$. More generally, if $A_1, A_2, \ldots$ is an infinite sequence of pairwise disjoint measurable sets, then

$$\mathrm{Vol}_d\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{Vol}_d(A_i).$$

Define a d-dimensional hyperplane in $\mathbb{R}^n$ to be a set obtained from a d-dimensional linear subspace by translation. For every d-dimensional hyperplane W in $\mathbb{R}^n$ we can let $w : W \to \mathbb{R}^d$ be any (Euclidean) distance-preserving bijection and define $\mathrm{Vol}_d(\cdot)$ on measurable subsets of W by specifying that $\mathrm{Vol}_d(S) = \mathrm{Vol}_d(w(S))$. This definition of $\mathrm{Vol}_d(S)$ doesn't depend on the choice of distance-preserving bijection, because Vol_d is invariant under translations and rotations.

The volumes of d-dimensional and $(d-1)$-dimensional sets are related by the following integral formula.

Fact 3.1. *If $S \subseteq \mathbb{R}^d$ is measurable and $W_s = \{\mathbf{x} \in \mathbb{R}^d \mid x_1 = s\}$, then*

$$\mathrm{Vol}_d(S) = \int_{-\infty}^{\infty} \mathrm{Vol}_{d-1}(S \cap W_s)\, ds.$$

Using Fact 3.1 we can derive the formula for the volume of a cone. If T is a subset of $\mathbb{R}^{d-1}$ and $h > 0$, then a *cone of height h with base T* is any set congruent to the following subset of $\mathbb{R}^d = \mathbb{R} \times \mathbb{R}^{d-1}$:

$$\mathrm{Cone}(T, h) = \{\mathbf{x} = ((1-t)h, t\mathbf{y}) \mid 0 \le t \le 1, \mathbf{y} \in T\}.$$

Fact 3.2. *If $T \subseteq \mathbb{R}^{d-1}$ is measurable, the volume of* $\mathrm{Cone}(T, h)$ *is* $\frac{h}{d} \mathrm{Vol}_{d-1}(T)$.

Proof. The intersection $\mathrm{Cone}(T, h) \cap W_s$ is empty unless $0 \le s \le h$, and then its $(d-1)$-dimensional volume is $t^d \cdot \mathrm{Vol}_{d-1}(T)$, where t is the solution to the equation $s = (1-t)h$; in other words, $t = 1 - \frac{s}{h}$. Using Fact 3.1 and the substitution $t = 1 - \frac{s}{h}$ we obtain

$$\mathrm{Vol}_d(\mathrm{Cone}(T, h)) = \int_0^h \left(1 - \tfrac{s}{h}\right)^d \cdot \mathrm{Vol}_{d-1}(T)\, ds = \mathrm{Vol}_{d-1}(T) \cdot \int_0^1 h\, t^d\, dt = \tfrac{h}{d} \mathrm{Vol}_{d-1}(T),$$

as claimed. □

Finally, in evaluating the volumes of high-dimensional sets it will be useful for us to be able to estimate the factorial function up to a constant factor. The following lemma furnishes the required estimate.

Lemma 3.3. *For any positive integer n,*

$$\sqrt{en}\left(\frac{n}{e}\right)^n < n! < e\sqrt{n}\left(\frac{n}{e}\right)^n. \tag{6}$$

Proof. Upon taking logarithms, the inequalities stated in the lemma become equivalent to

$$n\ln(n) - n + \tfrac{1}{2}\ln(n) + \tfrac{1}{2} < \ln(n!) < n\ln(n) - n + \tfrac{1}{2}\ln(n) + 1, \tag{7}$$

and we will prove the stated inequalities in this equivalent form. For all k and all $t \in (0,1)$ we have

$$\ln(k) + t\,(\ln(k+1) - \ln(k)) < \ln(k+t) < \ln(k) + \tfrac{t}{k},$$

where the left inequality is derived from the fact that the logarithm function is strictly concave, and the right inequality is derived from strict concavity along with the fact that the derivative of the natural logarithm at k is $\frac{1}{k}$. Integrating with respect to t and applying the substitution $x = k+t$, we find that

$$\ln(k) + \tfrac{1}{2}\,(\ln(k+1) - \ln(k)) < \int_k^{k+1} \ln(x)\,dx < \ln(k) + \tfrac{1}{2k}. \tag{8}$$

Now, summing over $k = 1, \ldots, n-1$,

$$\ln(n!) - \tfrac{1}{2}\ln(n) < \int_1^n \ln(x)\,dx < \ln(n!) - \ln(n) + \frac{1}{2}\sum_{k=1}^{n-1}\frac{1}{k} < \ln(n!) - \tfrac{1}{2}\ln(n) + \tfrac{1}{2}, \tag{9}$$

where we have used the fact that $\sum_{k=1}^{n-1}\frac{1}{k} < \ln(n) + 1$. (To derive that inequality, write the sum on the left as $1 + \sum_{k=2}^{n-1}\frac{1}{k}$ and note that this is bounded above by $1 + \int_1^{n-1}\frac{dt}{t}$.) Rearranging terms in (9) and using the fact that $\int_1^n \ln(x)\,dx = n\ln(n) - n + 1$, we derive

$$n\ln(n) - n + \tfrac{1}{2}\ln(n) + \tfrac{1}{2} < \ln(n!) < n\ln(n) - n + \tfrac{1}{2}\ln(n) + 1 \tag{10}$$

as claimed. □

3.2 Volume distribution near boundary

In this section we will explore a simple consequence of the rule for how Vol_d transforms under scaling, $\mathrm{Vol}_d(\lambda \cdot S) = \lambda^d \cdot \mathrm{Vol}_d(S)$. We'll see that this implies almost all of a high-dimensional sphere's volume is concentrated in a thin shell near the surface of the sphere.

Proposition 3.4. *Let $B^d(r)$ denote the Euclidean ball of radius r centered at $\mathbf{0} \in \mathbb{R}^d$, i.e. the ball of radius r in the L_2 norm. For any $c > 0$, the set of points whose distance from the boundary of $B^d(1)$ is greater than c/d constitutes less than e^{-c} fraction of the ball's volume.*

Proof. If $c \geq d$, then the set of points whose distance from the boundary of $B = B^d(1)$ is greater than c/d is empty. Otherwise, the set is equal to the interior of the ball $B^d(1 - \frac{c}{d})$, so its volume is equal to

$$\left(1 - \tfrac{c}{d}\right)^d \mathrm{Vol}_d(B).$$

To finish up, we use the inequality $1 - x < e^{-x}$ which is valid for all non-zero $x \in \mathbb{R}$. Applying this inequality with $x = \frac{c}{d}$, we find that

$$\left(1 - \tfrac{c}{d}\right)^d < \left(e^{-c/d}\right)^d = e^{-c}$$

which completes the proof of the proposition. □

3.3 Estimating the volume of the Euclidean ball

Let $B_1^d(r), B_2^d(r), B_\infty^d(r)$ denote the unit balls of radius r in $\mathbb{R}^d$ under the L_1, L_2, and L_∞ norms, respectively. In this section we will show that the volume of $B_2^d(1)$ is $d^{-d/2+o(d)}$, where the expression $o(d)$ in the exponent indicates an error term that grows sublinearly in d, as $d \to \infty$. To do so, we will inscribe an L_∞ ball inside $B = B_2^d(1)$ and circumscribe an L_1 ball around it, and we'll bound the volume of B from below and above by these inscribed and circumscribed shapes.

Lemma 3.5. *For any dimension $d \geq 1$, $B_\infty^d(d^{-1/2}) \subset B_2^d(1) \subset B_1^d(d^{1/2})$.*

Proof. Every $\mathbf{x} \in B_\infty^d(d^{-1/2})$ satisfies $|x_i| \leq d^{-1/2}$ for $i = 1, 2, \dots, d$, which implies

$$\sum_{i=1}^d x_i^2 \leq \sum_{i=1}^d \tfrac{1}{d} = 1,$$

hence $\mathbf{x} \in B_2^d(1)$. To prove $B_2^d(1) \subseteq B_1^d(d^{1/2})$, consider any $\mathbf{x} \in B_2^d(1)$ and let $\mathbf{y}$ denote a vector in $\{\pm 1\}^d$ such that $x_i y_i \geq 0$ for all i; in other words, $y_i = -1$ if $x_i < 0$, $y_i = 1$ if $x_i > 0$, and y_i is an arbitrary element of $\{\pm 1\}$ if $x_i = 0$. We have

$$\|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i| = \sum_{i=1}^d x_i y_i \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2,$$

where the last step is the Cauchy-Schwartz Inequality. Recalling that $\|\mathbf{x}\|_2 \leq 1$ and calculating that $\|\mathbf{y}\|_2 = d^{1/2}$, we find that $\|\mathbf{x}\|_1 \leq d^{1/2}$, as claimed. □

Lemma 3.6. *The unit balls of the L_1 and L_∞ norms $\mathbb{R}^d$ have volumes*

$$\mathrm{Vol}_d(B_1^d(1)) = \frac{2^d}{d!}, \quad \mathrm{Vol}_d(B_\infty^d(1)) = 2^d.$$

Proof. The L_∞ ball B_∞^d is simply the set $[-1, 1]^d$ of vectors whose coordinates are all between -1 and 1. This is a product of d intervals of length 2, so its volume is 2^d.

To estimate $\mathrm{Vol}_d(B_1^d)$, first dissect B_1^d into two congruent pieces: one consisting of the vectors in B_1^d whose first coordinate is non-negative, and the other consisting of the vectors in B_1^d whose first coordinate is non-positive. (These sets have a non-empty intersection consisting of vectors whose first coordinate is zero, but the d-dimensional volume of this intersection is zero.) Both pieces of this dissection are congruent to $\mathrm{Cone}(B^{d-1}, 1)$. Hence,

$$\mathrm{Vol}_d(B^d) = 2\,\mathrm{Vol}_d(\mathrm{Cone}(B^{d-1}, 1)) = \frac{2}{d}\,\mathrm{Vol}_{d-1}(B^{d-1}).$$

Solving this recurrence with the base case $\mathrm{Vol}_1(B^1) = 2$, we obtain $\mathrm{Vol}_d(B^d) = \frac{2^d}{d!}$. □

Proposition 3.7. *The volume of the Euclidean unit ball in $\mathbb{R}^d$ satisfies*

$$\left(\frac{2}{\sqrt{d}}\right)^d < \mathrm{Vol}_d(B_2^d(1)) < \left(\frac{2e}{\sqrt{d}}\right)^d. \tag{11}$$

Proof. By Lemma 3.5 we have $\mathrm{Vol}_d(B^d_\infty(d^{-1/2})) < \mathrm{Vol}_d(B^d_2(1)) < \mathrm{Vol}_d(B^d_1(d^{1/2}))$. Applying the rule that scaling a set in $\mathbb{R}^d$ scales its volume by λ to the formulas for $\mathrm{Vol}_d(B^d_\infty(1))$ and $\mathrm{Vol}_d(B^d_1(1))$, we can calculate the volumes of $B^d_\infty(d^{-1/2})$ and $B^d_1(d^{1/2})$ exactly and conclude that

$$2^d \cdot d^{-d/2} < \mathrm{Vol}_d(B^d_2(1)) < \frac{2^d \cdot d^{d/2}}{d!}. \tag{12}$$

From Lemma 3.3 we know that $\frac{1}{d!} < (\frac{e}{d})^d$, and substituting this upper bound for $\frac{1}{d!}$ into inequality (12), we obtain inequality (11). □

Above we have estimated the volume of a Euclidean unit ball by "sandwiching" it between the unit balls of the L_∞ and L_1 norms. A slightly more complicated way to obtain qualitatively similar estimates is to sandwich the d-dimensional ball between a cylinder and a cone. The benefit of the latter approach is that it enables us to estimate (within a constant factor) to volume ratio of the unit balls in d and $d-1$ dimensions, which will be helpful in the following section.

Lemma 3.8. *For any $\varepsilon > 0$, the Euclidean unit ball $B = B^d_2(1)$ is contained in the cone $C(\varepsilon) = \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, \varepsilon x_1 + \sqrt{(1-\varepsilon^2)(x_2^2 + x_3^2 + \cdots + x_d^2)} \le 1\right\}$.*

Proof. For any $\mathbf{x} \in B$, apply the Cauchy-Schwartz inequality to the two-dimensional vectors

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ \sqrt{x_2^2 + \cdots + x_d^2} \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} \varepsilon \\ \sqrt{1-\varepsilon^2} \end{bmatrix}.$$

Observe that $\|\mathbf{a}\|_2 = \|\mathbf{x}\|_2 \le 1$ since $\mathbf{x} \in B$, and that $\|\mathbf{b}\|_2 = 1$. Hence, the Cauchy-Schwartz Inequality implies $\langle \mathbf{a}, \mathbf{b} \rangle \le 1$. Expressing this inequality in terms of the coordinates of the vector $\mathbf{x}$, we find that $\mathbf{x}$ satisfies the inequality defining $C(\varepsilon)$. □

Lemma 3.9. *Let $B^d = B^d_2(1)$ and $B^{d-1} = B^{d-1}_2(1)$ denote the Euclidean unit balls in d and $d-1$ dimensions respectively. The volumes $\mathrm{Vol}_d(B^d)$ and $\mathrm{Vol}_{d-1}(B^{d-1})$ obey the following relation:*

$$\frac{2}{\sqrt{e}} \cdot \frac{\mathrm{Vol}_{d-1}(B^{d-1})}{\sqrt{d}} \le \mathrm{Vol}_d(B^d) \le 2\sqrt{e} \cdot \frac{\mathrm{Vol}_{d-1}(B^{d-1})}{\sqrt{d}}. \tag{13}$$

Proof. Let A denote the cylinder

$$A = \left\{\mathbf{x} \in \mathbb{R}^d \,\middle|\, x_1^2 \le \frac{1}{d},\; x_2^2 + x_3^2 + \cdots + x_d^2 \le \frac{d-1}{d}\right\}$$

and observe that $A \subset B^d$ since every $\mathbf{x} \in A$ satisfies $x_1^2 + x_2^2 + \cdots + x_d^2 \le \frac{1}{d} + \frac{d-1}{d} = 1$. The height of cylinder A is $\frac{2}{\sqrt{d}}$ and its base is a $(d-1)$-dimensional ball of radius

$$r = (1 + \frac{1}{d-1})^{-1/2} > e^{-1/(2d-2)}$$

so its volume is

$$\mathrm{Vol}_d(A) = \frac{2}{\sqrt{d}} r^{d-1}\, \mathrm{Vol}_{d-1}(B^{d-1}) > \frac{2}{\sqrt{e}} \cdot \frac{\mathrm{Vol}_{d-1}(B^{d-1})}{\sqrt{d}}. \tag{14}$$

Recall the infinite cone $C(\varepsilon)$ defined in Lemma 3.8, and let $-C(\varepsilon)$ denote the set $\{-\mathbf{x} \mid \mathbf{x} \in C(\varepsilon)\}$. The intersection $C = C(\varepsilon) \cap -C(\varepsilon)$ is a union two cones, each of height $\frac{1}{\varepsilon}$, whose common base is a $(d-1)$-dimensional ball whose radius is $(1-\varepsilon^2)^{-1/2}$. If we set $\varepsilon = \frac{1}{\sqrt{d}}$ then

$$(1-\varepsilon^2)^{-1/2} = \left(1 - \frac{1}{d}\right)^{-1/2} = \left(1 + \frac{1}{d-1}\right)^{1/2} < e^{1/2(d-1)}.$$

$$\mathrm{Vol}_d(C) = \frac{2}{d} \cdot \frac{1}{\varepsilon} \cdot \left(1-\varepsilon^2\right)^{-(d-1)/2} \cdot \mathrm{Vol}_{d-1}(B^{d-1}) < \frac{2\sqrt{e}}{\sqrt{d}} \cdot \mathrm{Vol}_{d-1}(B^{d-1}) \tag{15}$$

Since Lemma 3.8 tells us that $B^d \subset C(\varepsilon)$ and $B^d \subset -C(\varepsilon)$, we have $B \subset C$. Combining the set-theoretic relations $A \subseteq B \subseteq C$ with the volume bounds derived in Inequalities (14) and (15), we obtain the relation (13) asserted in the lemma statement. □

3.4 Volume distribution near equator

As one consequence of estimating the Euclidean ball's volume, we can prove that most of the volume is located in a thin layer near the equator. In fact, letting $B = B_2^d(1)$ denote the Euclidean unit ball in $\mathbb{R}^d$, if $L_i(w) = \{x \in B \mid -w \le x_i \le w\}$ denotes a layer of width $2w$ centered on the equatorial disc $\{x \in B \mid x_i = 0\}$, then we will prove that for any $c > 0$, the complement of $L_i = L_i\left(\sqrt{c/d}\right)$ contains only $2e^{-\beta c}$ fraction of the volume of B, for some constant $\beta > 0$.

As a warm-up before proving this exponentially small upper bound on the volume of $B \setminus L_i$, let us prove a simpler upper bound showing that for any $c > 1$, the set $C_i = B \setminus L_i(\sqrt{c/d})$ contains at most $\frac{1}{c}$ of the volume of B. The key observation is that every point $\mathbf{x} \in B$ belongs to fewer than d/c of the sets $C_1, C_2, \ldots, C_d$. Indeed, if $\mathbf{x} \in C_i$ then $x_i^2 > c/d$, and the constraint $\sum_{i=1}^d x_i^2 \le 1$ ensures that fewer than d/c indices i satisfy the inequality $x_i^2 > c/d$. Since $C_1, C_2, \ldots, C_d$ are subsets of B and every point of B belongs to fewer than d/c of them, their combined volume is less than $\frac{d}{c}\,\mathrm{Vol}(B)$. Since all of the sets are congruent to each other, they all have the same volume, which must therefore be less than $\frac{1}{c}\,\mathrm{Vol}(B)$.

Proposition 3.10. *Let $B = B_2^d(1)$ denote the Euclidean unit ball, and for some $c \ge 4$ let $L = L_i(\sqrt{c/d})$ denote the layer of width $2\sqrt{c/d}$ around the equator. The volume of $B \setminus L$ satisfies*

$$\mathrm{Vol}_d(B \setminus L) < \sqrt{\frac{e}{c}}\, e^{-c/2}\, \mathrm{Vol}_d(B).$$

Proof. If $c \ge d$ then $B \setminus L$ is an empty set and there is nothing to prove. Assume henceforth that $c < d$. Then the set $B \setminus L$ consists of two congruent spherical caps. The base of each spherical cap is a $(d-1)$-dimensional ball $B_2^{d-1}(r)$ whose radius r satisfies $r^2 + \frac{c}{d} = 1$. For $c \ge 4$ this implies $r < 1 - \frac{c}{2(d-1)} < e^{-c/2(d-1)}$. Applying Lemma 3.8 with $\varepsilon = \sqrt{\frac{c}{d}}$, we know that B is contained in the infinite cone $C(\varepsilon)$. The portion of this cone sitting above the hyperplane $\{x_1 = \varepsilon\}$ has base consisting of the points $\mathbf{x}$ such that $x_1 = \varepsilon$ and $x_2^2 + x_3^2 + \cdots + x_d^2 \le 1 - \varepsilon^2 = 1 - \frac{c}{d} = r^2$; this matches the base of the spherical cap. Hence, the volume of the spherical cap is less than the volume of the cone, which is

$$\frac{1}{d}\cdot\left(\frac{1}{\varepsilon}-\varepsilon\right)\cdot r^{d-1}\cdot \mathrm{Vol}_{d-1}(B^{d-1}) < \sqrt{\frac{1}{cd}}\cdot e^{-c/2}\cdot \mathrm{Vol}_{d-1}(B^{d-1}) \qquad < \frac{1}{2}\sqrt{\frac{e}{c}}e^{-c/2}\cdot \mathrm{Vol}_d(B^d),$$

where the last inequality follows from Lemma 3.9. □

3.5 Random high-dimensional vectors are nearly orthogonal

In d dimensions, if non-zero vectors $\mathbf{z}_1, \dots, \mathbf{z}_k$ are pairwise orthogonal, meaning that the dot product of any two of them is zero, then the vectors are linearly independent[2] and thus k must be less than or equal to d. In this section we will see that the situation is completely different if we require the vectors to be *nearly orthogonal*, meaning that the angle between any two of them lies in the interval from $\frac{\pi}{2}-\varepsilon$ to $\frac{\pi}{2}+\varepsilon$ radians, for some arbitrarily small $\varepsilon > 0$. We will prove that the maximum number of pairwise nearly orthogonal vectors in d dimensions grows exponentially with d, for any fixed $\varepsilon > 0$. The proof that we present shows, in fact, that if $m < e^{-\varepsilon^2 d/16}$, then with high probability a random set of m vectors sampled independently and uniformly at random from the unit ball B^d in d-dimensional Euclidean space will be pairwise nearly orthogonal. This is an illustration of a powerful technique called the *probabilistic method* in which one proves that an object having a certain property exists by showing that a random object possesses the property with positive probability. (In the case presented here, the "object" in question is a collection of m vectors in $\mathbb{R}^d$, and the property is pairwise near orthogonality.) In many cases, including this one, directly constructing an object with the required property is much more difficult than proving the existence of such an object using the probabilistic method.

At the heart of our proof that a random m-tuple of vectors in B^d are likely to be pairwise nearly orthogonal is the following lemma concerning the probability that two random vectors form an angle that differs from $\frac{\pi}{2}$ by more than ε.

Lemma 3.11. *Suppose* $\mathbf{x}, \mathbf{y}$ *are two random vectors sampled independently and uniformly at random from* B^d*. Let* $\theta \in [0,\pi]$ *denote the angle formed between* $\mathbf{x}$ *and* $\mathbf{y}$*. For any* ε *such that* $0 < \varepsilon < \frac{1}{8}$ *and any* $d > \frac{4e}{\varepsilon^2}$*, the probability that* $|\frac{\pi}{2}-\theta| > \varepsilon$ *is less than* $2e^{-\varepsilon^2 d/8}$.

Proof. The joint distribution of the pair $\mathbf{x}, \mathbf{y}$ is rotation-invariant, so the conditional distribution of the angle θ given that $\mathbf{y}$ is parallel to the standard basis vector $\mathbf{e}_1$ is the same as the unconditional distribution of θ. Furthermore, since θ depends only on the orientations of $\mathbf{x}$ and $\mathbf{y}$, not their lengths, we can condition on the event $\mathbf{y} = \mathbf{e}_1$ without affecting the distribution of θ.

Recalling now that the dot product of two vectors is equal to the product of their lengths, times the cosine of the angle between them, we find that our assumption that $\mathbf{y} = \mathbf{e}_1$ allows us to calculate the cosine of θ as follows.

$$\cos(\theta) = \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{||\mathbf{x}||_2||\mathbf{y}||_2} = \frac{x_1}{||\mathbf{x}||_2}. \tag{16}$$

[2] One way to see this must be the case is to consider the linear function $\mathbf{f}_i(\mathbf{x}) = \langle \mathbf{z}_i, \mathbf{x}\rangle$ for $i = 1, 2, \dots, k$. By assumption, $\mathbf{f}_($$\mathbf{x})$ evaluates to zero at $\mathbf{z}_j$ for any $j \neq i$, hence $\mathbf{f}(\mathbf{x}) = 0$ whenever $\mathbf{x}$ is a linear combination of $\{\mathbf{z}_j \mid j \neq i\}$. However, $\mathbf{f}(\mathbf{z}_i) = \langle \mathbf{z}_i, \mathbf{z}_i\rangle > 0$, so $\mathbf{z}_i$ is not a linear combination of $\{\mathbf{z}_j \mid j \neq i\}$. Since this holds for every i, we may conclude that $\mathbf{z}_1, \dots, \mathbf{z}_k$ are linearly independent as claimed.

Using the identity $\sin\left(\frac{\pi}{2} - \theta\right) = \cos(\theta)$, we find that the event $|\frac{\pi}{2} - \theta| > \varepsilon$ is equivalent to the event

$$|\sin(\theta)| > \sin(\varepsilon).$$

The inequality $\sin(\varepsilon) > \varepsilon/2$ is valid whenever $0 < \varepsilon < \frac{1}{8}$, hence

$$\Pr\left(\left|\frac{\pi}{2} - \theta\right| > \varepsilon\right) \leq \Pr\left(\frac{|x_1|}{\|\mathbf{x}\|_2} > \frac{\varepsilon}{2}\right) \leq \Pr\left(\|\mathbf{x}\|_2 < 1 - \varepsilon\right) + \Pr\left(|x_1| > (1 - \varepsilon) \cdot \frac{\varepsilon}{2}\right). \tag{17}$$

The second inequality follows because the inequality $\frac{|x_1|}{\|\mathbf{x}\|_2} > \frac{\varepsilon}{2}$ is only satisfied when at least one of the following two events happens: either $\|\mathbf{x}\|_2 < 1 - \varepsilon$ or $|x_1| > (1 - \varepsilon) \cdot \frac{\varepsilon}{2}$. Therefore the event $\frac{|x_1|}{\|\mathbf{x}\|_2} > \frac{\varepsilon}{2}$ is contained in the union of the latter two events, and its probability is bounded above by the sum of their probabilities.

Proposition 3.4 implies that the probability of the event $\|\mathbf{x}\|_2 < 1 - \varepsilon$ is less than $e^{-\varepsilon d}$, which is less than $e^{-\varepsilon^2 d/8}$ due to our assumption that $\varepsilon < 1/8$. Applying Proposition 3.10 with $c = \varepsilon^2 d/4$, and hence $\sqrt{c/d} = \varepsilon/2$, we find that the probability of the event $|x_1| > \varepsilon/2$ is less than $\sqrt{4e\varepsilon^2 d}e^{-\varepsilon^2 d/8}$, which is less than $e^{-\varepsilon^2 d/8}$ due to our assumption that $d > \frac{4e}{\varepsilon^2}$. To sum up, we have shown that both probabilities on the right side of (17) are less than $e^{-\varepsilon^2 d/8}$, hence the probability on the left side is less than $2e^{-\varepsilon^2 d/8}$. □

Proposition 3.12. *For every ε, d, m satisfying $0 < \varepsilon < \frac{1}{8}$, $d > \frac{4e}{\varepsilon^2}$, $m < e^{\varepsilon^2 d/16}$, if vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ are drawn independently and uniformly at random from the Euclidean unit ball $B^d \subset \mathbb{R}^d$, then with probability at least $\frac{1}{2m}$ every pair of vectors in the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ forms an angle between $\frac{\pi}{2} - \varepsilon$ and $\frac{\pi}{2} + \varepsilon$.*

Proof. We can use Lemma 3.11 to put an upper bound on the expected number of pairs $\mathbf{x}_i, \mathbf{x}_j$ that form an angle θ such that $|\frac{\pi}{2} - \theta| > \varepsilon$. The probability that any one such pair forms such an angle is less than $2e^{-\varepsilon^2 d/8}$, which is less than $\frac{2}{m^2}$ by our assumption on m. The number of unordered pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$ with $i \neq j$ is

$$\binom{m}{2} = \frac{m^2 - m}{2}.$$

By linearity of expectation, the expected number of (unordered) pairs $\{\mathbf{x}_i, \mathbf{x}_j\}$ that form an angle θ not lying between $\frac{\pi}{2} - \varepsilon$ and $\frac{\pi}{2} + \varepsilon$ is less than

$$\frac{2}{m^2} \cdot \frac{m^2 - m}{2} = 1 - \frac{1}{2m}.$$

The proposition follows, since a non-negative integer-valued random variable must always satisfy the inequality $\mathbb{E}[X] \geq \Pr(X > 0)$. □

4 Matrices

A matrix is a two-dimensional array of real numbers, M, with entries denoted by M_{ij}. Here, the ranges of i and j are finite intervals $[m] = \{1, 2, \ldots, m\}$ and $[n] = \{1, 2, \ldots, n\}$, where m and n are the number of *rows* and *columns*, respectively, of the matrix M.

Matrices play at least three distinct important roles in mathematics, computer science, and data science.

1. They encode information that takes the form of a two-dimensional array. A running example in this section will be a matrix encoding course enrollments in a department, with two rows that tabulate the number of undergraduate and graduate students, respectively, and with one column for each course offered by the department. In this example, the column for course j would contain entries M_{1j} and M_{2j} encoding the number of undergraduates and grad students, respectively, enrolled in course j.

2. An $m \times n$ matrix can encode a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$. In this encoding, the matrix M encodes the function $T : \mathbb{R}^n \to \mathbb{R}^m$ where $T(\mathbf{x})$ is the vector $\mathbf{y} \in \mathbb{R}^m$ whose coordinates are defined, for each $i \in [m]$ by the equation

$$y_i = \sum_{j=1}^{n} M_{ij} x_j.$$

3. An $m \times n$ matrix can also encode a *bilinear function* on $\mathbb{R}^m \times \mathbb{R}^n$. A function $A : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ is called bilinear if it satisfies

$$\begin{aligned} \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^m, \mathbf{z} \in \mathbb{R}^n \quad & A(a\mathbf{x} + b\mathbf{y}, \mathbf{z}) = aA(\mathbf{x}, \mathbf{z}) + bA(\mathbf{y}, \mathbf{z}) \\ \forall \mathbf{x} \in \mathbb{R}^m, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n \quad & A(\mathbf{x}, a\mathbf{y} + b\mathbf{z}) = aA(\mathbf{x}, \mathbf{y}) + bA(\mathbf{x}, \mathbf{z}). \end{aligned}$$

   Equivalently, A is bilinear if and only if for every $\mathbf{y} \in \mathbb{R}^n$ the function $f(\mathbf{x}) = A(\mathbf{x}, \mathbf{y})$ is a linear function of $\mathbf{x}$, and for every $\mathbf{x} \in \mathbb{R}^m$ the function $g(\mathbf{y}) = A(\mathbf{x}, \mathbf{y})$ is a linear function of $\mathbf{y}$. We say that matrix M encodes the bilinear function $A : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$ if

$$A(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} M_{ij} x_i y_j.$$

4.1 Change of basis

One of the tricky things about working with matrices is that we often want to write a matrix representing "the same thing" as M using a different basis. Doing this can be confusing because the way to rewrite M depends on what "thing" we are encoding using M.

Example 4.1. Let us return to our running example of a matrix M with 2 rows and n columns, representing the enrollments of n courses by noting the number of undergraduate students in the first row and the number of graduate students in the second row. A different matrix representing the same information might have the total number of students in the first row and the number of graduate students in the second row. Let us call this second matrix M'. Its relationship to M can be expressed by the formulas

$$M'_{1j} = M_{1j} + M_{2j}, \qquad M'_{2j} = M_{2j}$$

or more succinctly by the equation

$$M' = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] M.$$

The matrix $B = \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ is a "change of basis" matrix describing how the entries of M transform when we rewrite the data in the format of M'.

To illustrate the subtlety of working with change-of-basis matrices, let us now suppose that the university's budget model credits the department with \$2 for every undergraduate student and \$1 for every graduate student. (These aren't realistic numbers, we're just using them for the sake of this example.) Consider course j whose enrollment is represented in the first basis by the vector $\mathbf{m}_j = \left[\begin{smallmatrix} M_{1j} \\ M_{2j} \end{smallmatrix}\right]$ and in the second basis by the vector $\mathbf{m'}_j = \left[\begin{smallmatrix} M'_{1j} \\ M'_{2j} \end{smallmatrix}\right]$. The department's revenue from course j can be calculated by the expression $[\,2\;1\,]\mathbf{m}_j$ (\$2 for every undergraduate student plus \$1 for every graduate student), but it can also be calculated by the expression $[\,2\;{-1}\,]\mathbf{m'}_j$ (\$2 for every student, minus \$1 for every graduate student). Evidently, the change of basis which transforms $\mathbf{m}_j$ to $\mathbf{m'}_j$ also transforms the linear function represented by the row vector $[\,2\;1\,]$ to the one represented by the row vector $[\,2\;{-1}\,]$, even though

$$[\,2\;1\,] \cdot \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] \neq [\,2\;{-1}\,] \qquad \text{and} \qquad \left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right] \cdot \left[\begin{smallmatrix} 2 \\ 1 \end{smallmatrix}\right] \neq \left[\begin{smallmatrix} 2 \\ -1 \end{smallmatrix}\right].$$

What's going on here is that a linear function represented in the first basis by a row vector $\mathbf{r}$ becomes represented in the second basis by a row vector $\mathbf{r'} = \mathbf{r}B^{-1}$. A change of basis which operates on vectors via left multiplication by B operates on linear functions (represented as row vectors) via right multiplication by B^{-1}. If we choose to represent a linear function of $\mathbf{m}$ as an inner product $\langle \mathbf{c}, \mathbf{m} \rangle$, where $\mathbf{c}$ is a column vector, then the change-of-basis formula becomes even more obscure: the change of basis that transforms $\mathbf{m}$ to $B\mathbf{m}$ acts on $\mathbf{c}$ by transforming it into $(B^{-1})^{\mathsf{T}}\mathbf{c}$.

To derive the correct change-of-basis formulae for different types of vectors and matrices it is useful to introduce the notion of a *based vector space.* This is not a widely used mathematical term, but just a useful term we are using in this course to simplify the discussion of how to account for a change of basis.

Definition 4.1. A *based vector space* is a finite-dimensional vector space V together with a choice of isomorphism $\beta : \mathbb{R}^n \to V$ for some $n \in \mathbb{N}$.

Recall that $\mathbb{R}^n$ has a *standard basis* $\mathbf{e}_1, \ldots, \mathbf{e}_n$ where $\mathbf{e}_i$ has a 1 in its i^{th} coordinate and 0 in every other coordinate. If V is a based vector space then the vectors $\beta(\mathbf{e}_1), \ldots, \beta(\mathbf{e}_n)$ constitute a basis of V. Conversely, if $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is an ordered n-tuple of vectors that form a basis of V, then there is a unique isomorphism $\beta : \mathbb{R}^n \to V$ such that $\beta(\mathbf{e}_i) = \mathbf{v}_i$. Thus, giving a vector space V the structure of a based vector space is equivalent to choosing a basis for V and arranging the elements of the basis into an ordered n-tuple.

For a vector space V whose elements are semantically meaningful (e.g., course enrollments rather than abstract ordered pairs of numbers), giving V the structure of a based vector space is tantamount to settling on a convention for how to represent elements of V as n-tuples of numbers. This phenomenon already exists — and is well known — in the context of one-dimensional vector spaces, where the process of representing physical quantities as numbers requires choosing units. For example, it is meaningless to say, "The mass of my textbook is 2.5," whereas the sentence, "The mass of my textbook is 2.5 kilograms," is perfectly meaningful. In this case, masses of physical objects can be interpreted as elements of an abstract one-dimensional vector space V in which addition represents the operation of combining two disjoint physical objects. Two different choices of units, such as kilograms versus grams, are represented by two differed based vector space structures $\beta_{\text{kg}} : \mathbb{R} \to V$ and $\beta_{\text{g}} : \mathbb{R} \to V$ that send the element $1 \in \mathbb{R}$ to a one-kilogram mass and a one-gram mass, respectively. Choosing different based vector space structures for a higher-dimensional vector space V can be interpreted as a higher-dimensional counterpart to the process of converting between two different systems of units.

Definition 4.2. If V is an n-dimensional vector space and $\beta_1 : \mathbb{R}^n \to V$ and $\beta_2 : \mathbb{R}^n \to V$ are two different based vector space structures on V, the linear transformation $\beta_2^{-1} \circ \beta_1 : \mathbb{R}^n \to \mathbb{R}^n$ is represented by an $n \times n$ matrix called the *change of basis matrix* from β_1 to β_2.

Example 4.2. Returning to our running example, course enrollments can be interpreted as elements of an abstract two-dimensional vector space V. When course enrollments are represented as columns of the matrix M in Example 4.1, this corresponds to choosing a based vector space structure β_1 on V that sends $\mathbf{e}_1$ to the element of V represented in matrix M by the column vector $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, i.e. a course with one undergraduate and no graduate students. Let us denote this element of V by $\mathbf{u}$, for "undergraduate". Meanwhile $\beta_1(\mathbf{e}_2)$ is the element of V represented in matrix M by the column vector $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, i.e. a course with no undergraduates and one graduate student. Let us denote this element of V by $\mathbf{g}$, for "graduate".

The matrix M' represents course enrollments (i.e., elements of V) in an alternate data format that corresponds to a different based vector space structure. In this structure, $\beta_2(\mathbf{e}_1)$ is the element of V represented in matrix M by the column vector $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$, i.e. a course with one student in total, but zero graduate students. This is again the vector $\mathbf{u} \in V$. However, $\beta_2(\mathbf{e}_2)$ is the element of V represented in matrix M' by the column vector $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$, i.e. a course with *zero students in total, but one graduate student!* It's a bit hard to wrap one's head around what this means, but the most natural way to interpret it is that adding this vector to a course enrollment represents the operation of one undergraduate dropping the course and being replaced by a graduate student. (That operation has zero effect on the total number of students, but it increments the number of graduate students.) In other words, $\beta_2(\mathbf{e}_2) = \mathbf{g} - \mathbf{u}$.

Now, let's compute the change of basis matrix B. It is a two-by-two matrix whose columns are $B\mathbf{e}_1$ and $B\mathbf{e}_2$. We can calculate each column as follows.

$$\begin{aligned} B\mathbf{e}_1 &= \beta_2^{-1}(\beta_1(\mathbf{e}_1)) = \beta_2^{-1}(\mathbf{u}) = \beta_2^{-1}(\beta_2(\mathbf{e}_1)) = \mathbf{e}_1 \\ B\mathbf{e}_2 &= \beta_2^{-1}(\beta_1(\mathbf{e}_2)) = \beta_2^{-1}(\mathbf{g}) = \beta_2^{-1}(\mathbf{u} + (\mathbf{g} - \mathbf{u})) = \beta_2^{-1}(\beta_2(\mathbf{e}_1) + \beta_2(\mathbf{e}_2)) = \mathbf{e}_1 + \mathbf{e}_2. \end{aligned}$$

Hence, $B = \left[\begin{smallmatrix}1&1\\0&1\end{smallmatrix}\right]$, consistent with the change of basis formula for converting matrix M to M' derived in Example 4.1.

As the preceding example makes clear, when a matrix M represents a data table whose columns belong to a vector space V, if we change from one basis of V to another, the matrix M is transformed to BM, where B is the change-of-basis matrix. When M represents a linear transformation from V to W or a bilinear function on $V \times W$, the rules for how M transforms under a change of basis for V or W (or both) can be derived by reasoning about the equations that must be satisfied after the change of basis.

For example, suppose V and W are vector spaces of dimensions n and m, respectively. Suppose V and W each have two different bases, represented by based vector space structures β_{V1} and β_{V2} in the case of V, and by β_{W1} and β_{W2} in the case of W. Let B_V and B_W denote the respective change of basis matrices. Consider any linear transformation $T : V \to W$ and let M_1, M_2 be the matrices that represent T with respect to the based vector space structures β_{V1}, β_{W1} and β_{V2}, β_{W2}, respectively. Then for all $\mathbf{x} \in V$,

$$T(\mathbf{x}) = \beta_{W1}(M_1(\beta_{V1}^{-1}(\mathbf{x})))$$
$$T(\mathbf{x}) = \beta_{W2}(M_2(\beta_{V2}^{-1}(\mathbf{x})))$$

hence

$$\beta_{W1} \circ M_1 \circ \beta_{V1}^{-1} = \beta_{W2} \circ M_2 \circ \beta_{V2}^{-1}.$$

To isolate a formula for M_2 we multiply the last equation on the left by β_{W2}^{-1} and on the right by β_{V2}, obtaining

$$M_2 = \beta_{W2}^{-1} \circ \beta_{W1} \circ M_1 \circ \beta_{V1}^{-1} \circ \beta_{V2} = B_W M_1 B_V^{-1}. \tag{18}$$

Example 4.3. We return once more to our running example of course enrollments. Recall that in Example 4.1, if V is the vector space of course enrollments, we defined a linear function $V \to \mathbb{R}$ represented in basis β_{V1} by the row vector $\mathbf{r} = [\,2\ 1\,]$. When we transform to the basis β_{V2}, the change-of-basis formula for a linear transformation specifies that we should transform $\mathbf{r}$ to $\mathbf{r}B_V^{-1}$. This explains the reason why right-multiplication by the inverse of the change-of-basis matrix is the appropriate way to transform the coefficient vector of a linear function.

Now let us explore how the matrix representing a bilinear function transforms under change of basis. Recall that a bilinear function A on $\mathbb{R}^m \times \mathbb{R}^n$ is represented by a matrix M satisfying

$$A(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \sum_{j=1}^{n} M_{ij} x_i y_j = \sum_{i=1}^{m} x_i \sum_{j=1}^{n} M_{ij} y_j = \langle \mathbf{x}, M\mathbf{y} \rangle.$$

More generally, if $A : V \times W \to \mathbb{R}$ is a bilinear function and β_V, β_W are based vector space structures on V and W, respectively, then the matrix M representing A with respect to these bases satisfies

$$\forall \mathbf{v} \in V,\ \mathbf{w} in W\ A(\mathbf{v}, \mathbf{w}) = \left\langle \beta_V^{-1}(\mathbf{v}), M\beta_W^{-1}(\mathbf{w}) \right\rangle.$$

As before, if V and W each have two based vector space structures, denoted by β_{V1}, β_{V2} and β_{W1}, β_{W2}, and the bilinear function A is represented by matrices M_1 and M_2 with respect to these two pairs of based vector space structures, then we have the equation

$$\forall \mathbf{v} in V,\ \mathbf{w} \in W\ \left\langle \beta_{V2}^{-1}(\mathbf{v}), M_2\beta_{W2}^{-1}(\mathbf{w}) \right\rangle = \left\langle \beta_{V1}^{-1}(\mathbf{v}), M_1\beta_{W1}^{-1}(\mathbf{w}) \right\rangle.$$

Let $\mathbf{v} = \beta_{V2}(\mathbf{x})$ and $\mathbf{w} = \beta_{W2}(\mathbf{y})$.

$$\forall \mathbf{x} \in \mathbb{R}^m,\ \mathbf{y} \in \mathbb{R}^n\ \langle \mathbf{x}, M_2\mathbf{y} \rangle = \left\langle \beta_{V1}^{-1}(\beta_{V2}(\mathbf{x})), \beta_{W1}^{-1}(\beta_{W2}(\mathbf{y})) \right\rangle = \left\langle B_V^{-1}\mathbf{x}, B_w^{-1}\mathbf{y} \right\rangle = \left\langle \mathbf{x}, (B_V^{-1})^\top M_1 B_W^{-1}\mathbf{y} \right\rangle$$

where the last step used the identity $\langle M\mathbf{x}, \mathbf{y} \rangle = \left\langle \mathbf{x}, M^\top \mathbf{y} \right\rangle$. Hence $M_2 = (B_V^{-1})^\top M_1 B_W^{-1}$.

4.2 Adjoints and orthogonality

Taking the transpose of a matrix is an important operation in linear algebra. When the matrix represents a linear transformation between abstract vector spaces, the linear transformation that corresponds to the transpose of the matrix is called its *adjoint* and is defined in the following lemma. Before stating the lemma, we need the following definition.

Definition 4.3. If V is a finite-dimensional vector space with an inner product $\langle \cdot, \cdot \rangle_V$, a based vector space structure $\beta : \mathbb{R}^n \to V$ is said to be *compatible* with the inner product structure on V if it satisfies

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \;\; \langle \mathbf{x}, \mathbf{y} \rangle = \langle \beta(\mathbf{x}), \beta(\mathbf{y}) \rangle_V$$

where $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the dot product of $\mathbf{x}$ and $\mathbf{y}$, i.e. the standard inner product structure on $\mathbb{R}^n$.

Lemma 4.1. *If V, W are finite-dimensional vector spaces, each equipped with a non-degenerate inner product, and $T : V \to W$ is a linear transformation, then there is a unique linear transformation $U : W \to V$ called the* adjoint *of T, that satisfies*

$$\forall \mathbf{v} \in V,\ \mathbf{w} \in W \;\; \langle T\mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{v}, U\mathbf{w} \rangle .$$

If β_V, β_W are based vector space structures on V, W that are compatible with their respective inner products, and M_T, M_U are the matrices representing T and its adjoint U, respectively, then M_U is the transpose of M_T.

Proof. Because the inner product structures on V and W are non-degenerate, there are isomorphisms $\iota_V : V \to V^*$ and $\iota_W : W \to W^*$ such that $\iota_V(\mathbf{x})$ is the linear function $\mathbf{f}$ that, when applied to a vector $\mathbf{y} \in V$, yields the inner product $\mathbf{f}(\mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$, and ι_W is defined similarly using the inner product on W. Let $T^* : W^* \to V^*$ denote the linear transformation such that for all $\mathbf{g} \in W^*$, $T^*(\mathbf{g})$ is the linear function $\mathbf{f} \in V^*$ defined by $\mathbf{f}(\mathbf{y}) = \mathbf{g}(T(\mathbf{y}))$. Let $U = \iota_V^{-1} \circ T^* \circ \iota_W$. Then for any $\mathbf{v} \in V, \mathbf{w} \in W$, if we let $\mathbf{f} = T^*(\iota_W(\mathbf{w}))$, then

$$\langle \mathbf{v}, U\mathbf{w} \rangle = \left\langle \mathbf{v}, \iota_V^{-1}(T^*(\iota_W(\mathbf{w}))) \right\rangle = \left\langle \mathbf{v}, \iota_V^{-1}(\mathbf{f}) \right\rangle = \left\langle \iota_V^{-1}(\mathbf{f}), \mathbf{v} \right\rangle = \mathbf{f}(\mathbf{v}) = \iota_W(\mathbf{w})(T\mathbf{v}) = \langle \mathbf{w}, T\mathbf{v} \rangle = \langle T\mathbf{v}, \mathbf{w} \rangle ,$$

which verifies that U satisfies the equation defining the adjoint of T. To verify that U is unique, observe that if U' also satisfies the defining equation of the adjoint, then for all $\mathbf{v} \in V,\ \mathbf{w} \in W$,

$$\langle \mathbf{v}, U\mathbf{w} - U'\mathbf{w} \rangle = \langle \mathbf{v}, U\mathbf{w} \rangle - \langle \mathbf{v}, U'\mathbf{w} \rangle = \langle T\mathbf{v}, \mathbf{w} \rangle - \langle T\mathbf{v}, \mathbf{w} \rangle = 0.$$

Since $\mathbf{v}$ was an arbitrary vector in V and the inner product on V is non-degenerate, this implies that $U\mathbf{w} - U'\mathbf{w} = 0$. Since $\mathbf{w}$ was an arbitrary vector in W, this means $U = U'$.

Finally, the fact that M_U is the transpose of M_T can be checked by verifying that the standard inner product on $\mathbb{R}^n$ satisfies $\langle M\mathbf{x}, \mathbf{y} \rangle = \left\langle \mathbf{x}, M^\top \mathbf{y} \right\rangle$ for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and matrices $M \in \mathbb{R}^{n \times n}$. □

A matrix $M \in \mathbb{R}^{n \times n}$ is called *symmetric* if $M = M^\top$, and it is called *orthogonal* if $M^\top$ is the inverse of M. Based on Lemma 4.1 we can generalize the definitions of symmetric and orthogonal matrices to the setting of abstract inner product spaces as follows.

Definition 4.4. If V is a vector space with a non-degenerate inner product and $T : V \to V$ is a linear transformation, we say that T is *self-adjoint* with respect to the inner product on V if it equal to its own adjoint. In other words, a self-adjoint linear transformation is one that satisfies the equation

$$\langle T\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, T\mathbf{y}\rangle$$

for all $\mathbf{x},\ \mathbf{y} \in V$. We say that T is *orthogonal* with respect to the inner product on V if its adjoint is T^{-1}. Equivalently, an orthogonal linear transformation is one that satisfies the equation

$$\langle T\mathbf{x}, T\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$$

for all $\mathbf{x},\ \mathbf{y} \in V$.

4.3 Symmetric positive definite matrices

A very important set of square matrices are the symmetric positive definite matrices, i.e. the set of all matrices that represent positive definite inner products on $\mathbb{R}^n$. There are a number of equivalent characterizations of symmetric positive definite matrices, and all of them are important in different contexts. In this section we present several equivalent characterizations and prove their equivalence. A key starting point for the proof is the following observation.

Lemma 4.2. *If V in a vector space of dimension $n < \infty$ with a positive definite inner product $\langle \cdot, \cdot \rangle_V$, then V is isomorphic to $\mathbb{R}^n$ with the standard inner product structure. In other words there is a based vector space structure $\beta : \mathbb{R}^n \to V$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,*

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \;\; \langle \mathbf{x}, \mathbf{y}\rangle = \langle \beta\mathbf{x}, \beta\mathbf{y}\rangle_V \,. \tag{19}$$

Proof. The proof is by induction on n. When $n = 0$ there is nothing to prove, since V and $\mathbb{R}^n$ are both singleton sets consisting of the vector $\mathbf{0}$, whose inner product with itself is 0.

For $n > 0$, let W be an $(n-1)$-dimensional subspace of V, equipped with the inner product structure obtained by restricting $\langle \cdot, \cdot \rangle_V$ to pairs of vectors in W. There is a linear transformation $T : V \to W^*$ that maps each vector $\mathbf{x} \in V$ to the linear function $f_{\mathbf{x}} : W \to \mathbb{R}$ defined by $f_{\mathbf{w}}(\mathbf{w}) = \langle \mathbf{x}, \mathbf{w}\rangle_V$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be a basis of V. The vectors $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n) \in W^*$ must be linearly dependent, since $\dim(W^*) = \dim(W) = n-1$. Hence we can express $\mathbf{0} \in W^*$ as a non-trivial linear combination

$$\mathbf{0} = \sum_{i=1}^{n} a_i T(\mathbf{v}_i) = T\left(\sum_{i=1}^{n} a_i \mathbf{v}_i\right)$$

where the coefficients $a_1, \ldots, a_n$ are not all equal to zero. Let $\mathbf{v} = \sum_{i=1}^{n} a_i \mathbf{v}_i$, which is a nonzero vector in V since $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is a basis and $a_1, \ldots, a_n$ are not all zero. Recalling the definition of the linear transformation T, we see that the equation $T(\mathbf{v}) = 0$ means

$$\forall \mathbf{w} \in W \;\; \langle \mathbf{v}, \mathbf{w}\rangle_V = 0. \tag{20}$$

Since the inner product on V is positive definite and $\mathbf{v} \neq \mathbf{0}$, we know that $\langle \mathbf{v}, \mathbf{v}\rangle_V > 0$. Rescaling $\mathbf{v}$ if necessary, we can assume $\langle \mathbf{v}, \mathbf{v}\rangle = 1$. The rescaling doesn't affect the validity of (20).

The induction hypothesis implies there is an isomorphism $\beta_W : \mathbb{R}^{n-1} \to W$ such that $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \beta_W \mathbf{x}, \beta_W \mathbf{y} \rangle_V$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{n-1}$. Let us now define $\beta : \mathbb{R}^n \to V$ by specifying that

$$\beta \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \beta_W \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} + x_n \mathbf{v}.$$

We must verify that this β satisfies Equation (19). For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, let $\mathbf{x}', \mathbf{y}'$ denote the vectors in $\mathbb{R}^{n-1}$ obtained by extracting the first $n-1$ coordinates of $\mathbf{x}$ and $\mathbf{y}$, respectively. We have

$$\begin{aligned} \langle \beta \mathbf{x}, \beta \mathbf{y} \rangle_V &= \langle \beta_W \mathbf{x}' + x_n \mathbf{v}, \beta_W \mathbf{y}' + y_n \mathbf{v} \rangle_V \\ &= \langle \beta_W \mathbf{x}', \beta_W \mathbf{y}' \rangle_V + x_n \langle \mathbf{v}, \beta_W \mathbf{y}' \rangle_V + y_n \langle \beta_W \mathbf{x}', \mathbf{v} \rangle_V + x_n y_n \langle \mathbf{v}, \mathbf{v} \rangle_V \end{aligned}$$

Thinking about the four terms on the right side, the induction hypothesis implies that the first term equals $\langle \mathbf{x}', \mathbf{y}' \rangle$, the second and third terms vanish because of equation (20), and the four term equals $x_n y_n$ because we normalized $\mathbf{v}$ to ensure $\langle \mathbf{v}, \mathbf{v} \rangle = 1$. Hence,

$$\langle \beta \mathbf{x}, \beta \mathbf{y} \rangle_V = \langle \mathbf{x}', \mathbf{y}' \rangle + x_n y_n = \langle \mathbf{x}, \mathbf{y} \rangle ,$$

as desired. □

Proposition 4.3. *For a square matrix $M \in \mathbb{R}^{n \times n}$ the following properties are equivalent.*

1. *The bilinear function $f(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, M\mathbf{y} \rangle$ is a positive definite inner product.*
2. *$M = BB^\mathsf{T}$ for some invertible square matrix B.*
3. *$M = BB^\mathsf{T}$ for some (possibly rectangular) matrix B whose column space is $\mathbb{R}^n$.*
4. *$M = \sum_{i=1}^m a_i \mathbf{x}_i \mathbf{x}_i^\mathsf{T}$ for some coefficients $a_1, \ldots, a_m > 0$ and some sequence of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ that contains a basis for $\mathbb{R}^n$.*
5. *$M = QDQ^\mathsf{T}$ for some orthogonal matrix Q and diagonal matrix D with positive diagonal entries.*
6. *M is a symmetric matrix whose eigenvalues are all strictly positive.*