Note 21. The Hilbert space of random variables

Alex Fu

Fall 2022

1 Introduction

Real-valued random variables $X\colon \Omega \to \mathbb{R}$ are functions. Collections of functions naturally form *function spaces*, which often inherit certain structures on the domain or codomain. For instance, functions to a vector space V themselves form a vector space, under the operations

$$(f+g)(x) := f(x) + g(x)$$
$$(cf)(x) := cf(x).$$

We have also seen how functions can converge to a limit *pointwise* when convergence makes sense in the codomain. For another example, the norm of a linear mapping can be defined as

$$\|f\|_{\mathsf{op}} := \sup_{x \neq 0} \frac{\|f(x)\|_Y}{\|x\|_X}$$

where $f\colon X \to Y$ maps between two normed vector spaces. Other properties of the domain and codomain we can consider include continuity, topology, boundedness, metric, etc.

Here, we will use the fact that $\mathbb{R}$ has sums, scalar multiples, products, and limits to develop the *Hilbert space of random variables* $\mathcal{H} := L^2(\Omega; \mathbb{R})$, which is equipped with an inner product $\langle X, Y \rangle := \mathbb{E}(XY)$. We will see how we can leverage geometric intuition for random variables, then find a solution to the motivating problem of *linear least squares error estimation*.

2 Preliminaries

The prototypical example to keep in mind is *n-dimensional Euclidean space* $\mathbb{R}^n$. Even if there are key distinctions between finite-dimensional and infinite-dimensional vector spaces, the geometric intuition of "arrows in space" may still prove useful.

2.1 Linear algebra

Definition 1 (Real vector space).

A **real vector space** is a set of *vectors* equipped with two operations, *vector addition* and *scalar multiplication*, where elements in $\mathbb{R}$ are called *scalars*. The properties a vector space satisfy are listed below, though for your reference only.

- Vector addition $v + w$ is *closed*, *associative*, has an *identity* element (the zero vector), and associates an *inverse* element $-v$ to each vector v.
- Scalar multiplication $c \cdot v$ is *closed*, *compatible* with the multiplication of real numbers, compatible with *unity* (the scalar 1), and *distributive* over vector addition.

Every finite-dimensional real vector space is *isomorphic*, or equivalent up to relabelling the vectors, to the Euclidean space $(\mathbb{R}^n, +, \cdot)$ for some n. These vectors can thus be represented as tuples $v = (x_1, \ldots, x_n)$. However, function spaces are often infinite-dimensional, which we define below.

Definition 2 (Linear combination; span; linear independence).

- A **linear combination** v of $v_1, \ldots, v_n$ is any result of addition and scalar multiplication applied to those vectors:

$$v = \sum_{i=1}^{n} c_i v_i = c_1 v_1 + \cdots + c_n v_n.$$

  An equation involving linear combinations is a *linear relation*, or less commonly a "linear dependence," among the constituent vectors.

- The **span** of a set of vectors S is the set of all possible linear combinations that can be formed from vectors in S:

$$\text{span}(S) := \{c_1 v_1 + \cdots + c_n v_n : c_i \in \mathbb{R},\ v_i \in S,\ n \in \mathbb{N},\ i = 1, \ldots, n\}.$$

  The span of S is also the minimal vector *subspace* that contains S, the set of "all vectors that S can reach." A subspace is simply a subset of V that is a vector space itself.

- A set of vectors S is **linearly independent** if there is no nontrivial linear relation among them, i.e. if there is no equation of the form

$$c_1 v_1 + \cdots + c_n v_n = 0$$

  for some nonzero c_i, so that no vector is "redundant information."

Definition 3 (Basis; dimension).

- A **basis** is a set of vectors that spans the whole space and is linearly independent. Every vector can be written as a *unique* linear combination of basis vectors.
- The **dimension** of a vector space is the size of any basis, which is well-defined as every basis has the same size!

2.2 Inner product spaces

Definition 4 (Real inner product space).

A real vector space may be equipped with an *inner product* $\langle \cdot, \cdot \rangle : V \times V \to [0, \infty)$ to become an **inner product space**. For any $u, v, w \in V$ and $c, d \in \mathbb{R}$, the inner product satisfies

1. **Positive definiteness.** $\langle v, v \rangle \geq 0$, with $\langle v, v \rangle = 0$ iff $v = 0$.
2. **Bilinearity.** $\langle cu + dv, w \rangle = c \langle u, w \rangle + d \langle v, w \rangle$.
3. **Symmetry.** $\langle u, v \rangle = \langle v, u \rangle$.

The motivating example of the inner product is the *dot product* in Euclidean space $\mathbb{R}^n$,

$$u \cdot v = u^{\mathsf{T}} v = \sum_{i=1}^{n} u_i v_i.$$

From the inner product, we can derive the ideas of *length*, *distance*, *angle*, and *orthogonality* for general inner product spaces, some of which might not seem geometric in nature.

Definition 5 (Norm; metric; angle; orthogonality).

The **norm**, *length*, or *magnitude* of a vector v is $\|v\| = \sqrt{\langle v, v \rangle}$, where the following hold.

1. **Positive definiteness.** $\|v\| \geq 0$, with $\|v\| = 0$ iff $v = 0$.
2. **Homogeneity.** $\|cv\| = |c| \cdot \|v\|$.
3. **Triangle inequality.** $\|v + w\| \leq \|v\| + \|w\|$.

The **metric** or *distance* between two vectors v and w is induced as $d(v, w) = \|v - w\|$. We can also define the **angle** θ between two vectors from the equation

$$\langle u, v \rangle = \|u\| \, \|v\| \cos \theta.$$

Finally, two vectors u and v are **orthogonal**, denoted $u \perp v$, if $\langle u, v \rangle = 0$.

The inner product space we will consider also has the property of *completeness**: if the distances between vectors in a sequence become arbitrarily small, then the sequence indeed converges. Such an inner product space is called a **Hilbert space**.

Definition 6 (Orthogonal set; unit vector; orthonormal basis).

A set of vectors is **orthogonal** if $\langle v_i, v_j \rangle = 0$ whenever $i \neq j$. A vector is a **unit vector** if its norm is 1. An **orthonormal** set of vectors is an orthogonal set of unit vectors.

Proposition 1 (Orthogonality implies linear independence).

An orthogonal set of vectors is linearly independent, but not necessarily the converse. So, an orthogonal set that spans the whole space must also be an orthogonal *basis*.

Proof. Suppose that $c_1 v_1 + \cdots c_n v_n = 0$, and for any $j = 1, \ldots, n$ consider the inner product

$$0 = \left\langle \sum_{i=1}^{n} c_i v_i, v_j \right\rangle = \sum_{i=1}^{n} c_i \langle v_i, v_j \rangle = c_j \langle v_j, v_j \rangle .$$

The zero vector cannot belong to an orthogonal set of vectors, so v_j must be nonzero. Then the coefficients c_j must be identically zero, which proves linear independence. For a counterexample to the converse, consider the vectors $(1, 0)$ and $(1, 1)$ in $\mathbb{R}^2$. $\square$

Proposition 2 (Properties of orthonormal bases).

Let $\{v_1, \ldots, v_n\}$ be an orthonormal basis of V. Then the following hold.

a. **Unitariness.** $\langle v_i, v_j \rangle = \mathbb{1}_{i=j}$ for every i and j.

b. **Unique basis representation**. For any $w \in V$,

$$w = \sum_{i=1}^{n} \langle w, v_i \rangle v_i, \qquad \|w\| = \sum_{i=1}^{n} \langle w, v_i \rangle^2 .$$

c. **Orthogonal decomposition**. For any $w \in V$, and for any partition of the basis vectors such that $U = \text{span}\{v_1, \ldots, v_k\}$ and $U^\perp := \text{span}\{v_{k+1}, \ldots, v_n\}$,

$$w = (w|_U) + (w|_{U^\perp}) := \sum_{i=1}^{k} \langle w, v_i \rangle v_i + \sum_{i=k+1}^{n} \langle w, v_i \rangle v_i,$$

where the two components of the sum $w|_U$ and $w|_{U^\perp}$ are orthogonal to each other.

Proposition 3 (Gram–Schmidt procedure).

From any basis $\{v_1, \ldots, v_n\}$, we can find an *orthogonal* basis $\{e_1, \ldots, e_n\}$ as follows.

1. Set $e_1 := v_1$.

i. For each $i = 2, \ldots, n$, set $e_i := v_i - \sum_{j=1}^{i-1} \langle v_i, e_j \rangle e_j$.

We can *normalize*, divide by the norm to result in a unit vector, at each step or at the end, to obtain an orthonormal basis. If we also discard redundant, or linearly dependent, vectors, we can turn any *spanning set* into an orthonormal basis.

A key invariant of the procedure is that the span is preserved at every step: $\text{span}\{v_1, \ldots, v_i\} = \text{span}\{e_1, \ldots, e_i\}$ for every $i = 1, \ldots, n$.

Optionally, compare the Gram–Schmidt procedure to the disjointization of a countable union or the chain rule for entropy. At each step, some degree of *redundancy* is removed: the linear combination in $v_i - \sum_{j=1}^{i-1} \langle v_i, e_j \rangle e_j$, the union in $A_n \setminus \bigcup_{i=1}^{n-1} A_n$, or the mutual information in $H(X_n) - I(X_n; X_1, \ldots, X_{n-1})$. The now-"independent" components share a similar invariant:

$$\text{span}\{e_1, \ldots, e_i\}_\perp = \text{span}\{v_1, \ldots, v_i\}$$

$$\bigcup_{i=1}^{n} B_i = \bigsqcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i$$

$$\sum_{i=1}^{k} H(X_i \mid X_1, \ldots, X_{i-1}) = H(X_1, \ldots, X_k).$$

This same motif will become important in finding the *innovation* for LLSE. By Proposition 2, orthogonalization is incredibly helpful in decomposing a vector into individual components which can be simply summed. We present a key example in the following subsection.

2.3 Orthogonal projections

Definition 7 (Orthogonal projection).

The **orthogonal projection** of a vector v onto a subspace U is the vector in U that minimizes the distance from v to any vector in U:

$$\text{proj}_U(v) := \underset{u \in U}{\operatorname{argmin}} \|v - u\|^2 .$$

For example, sunlight casts three-dimensional objects onto the two-dimensional surface of the ground as shadows. The orthogonal projection of a street light pole is the tip of its shadow when cast by the sun directly overhead: the tip of the shadow is closest to the tip of the pole.

Proposition 4 (Properties of orthogonal projections).

a. The orthogonal projection is unique.

b. Let $\{u_1, \ldots, u_n\}$ be an orthonormal basis for U. Then $\text{proj}_U(v) = \sum_{i=1}^{n} \langle v, u_i \rangle u_i$.

c. A vector w is equal to $\text{proj}_U(v)$ if and only if $w \in U$ and $v - w \perp u$ for all $u \in U$.

The proofs are left as exercises. As a hint, consider the following equation, where $\|\cdot\|^2 = \langle \cdot, \cdot \rangle$:

$$\|v - u\|^2 = \|(v - \text{proj}_U(v)) + (\text{proj}_U(v) - u)\|^2 .$$

We also invite you to try your hand at the following exercises.

- The zero vector 0 is orthogonal to every vector, and is the unique vector orthogonal to itself.
- **Pythagorean theorem**. If u and v are orthogonal, then $\|u + v\|^2 = \|u\|^2 + \|v\|^2$.

- **Cauchy-Schwarz inequality**. $|\langle u, v\rangle| \leq \|u\| \, \|v\|$. Equivalently, $\langle u, v\rangle^2 \leq \langle u, u\rangle \langle v, v\rangle$.
- **Parallelogram equality**. $\|u+v\|^2 + \|u-v\|^2 = 2(\|u\|^2 + \|v\|^2)$.
- The projection operator proj_U is a *linear transformation*.

3 Definition and properties

The majority of the work was done in the preliminary section. We only need to apply the machinery to the random variables $\{X : \Omega \to \mathbb{R}\}$, and we are free to wander the probabilistic playground.

1. First, we can define sums $X+Y$, scalar multiples cX, and products XY of random variables, which makes the function space $\{X : \Omega \to \mathbb{R}\}$ a vector space.
2. We may then notice that covariance, which is bilinear and symmetric, is a good candidate for an inner product. However, covariance is not positive definite: $\text{cov}(X, X) = \text{var}(X) = 0$ only implies that X is almost surely constant, not surely zero.

   However, if we pretend that every random variable is zero-mean, then $\text{cov}(X, Y) = \mathbb{E}(XY)$ *is* positive definite. It turns out that in general, $\langle X, Y\rangle = \mathbb{E}(XY)$ defines an inner product: it is positive definite, as $\mathbb{E}(X^2)$ implies $X \overset{\text{a.s.}}{=} 0$.
3. It is still possible that the inner product is not always defined. By the Cauchy-Schwarz inequality *for expectation*,
   $$\langle X, Y\rangle = \mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)} = \|X\| \, \|Y\| \, .$$
   So, if we want every inner product $\mathbb{E}(XY)$ to remain finite, we must require that $\|X\| < \infty$. Equivalently, we need every random variable to have finite second moment:
   $$\mathcal{L}^2(\Omega; \mathbb{R}) := \left\{X : \Omega \to \mathbb{R} \mid \mathbb{E}(X^2) < \infty\right\}.$$
4. The final point is a bit subtle: the "inner product" is not truly positive definite unless $X \overset{\text{a.s.}}{=} 0$ implies $X = 0$. So, we take two random variables to be equal if they are almost surely equal, which turns out also guarantees the completeness of our space.
   $$L^2(\Omega; \mathbb{R}) := \mathcal{L}^2(\Omega; \mathbb{R}) \, / \overset{\text{a.s.}}{=}.$$

Definition 8 (Hilbert space of random variables).

The **Hilbert space of random variables** is $\mathcal{H} := L^2(\Omega; \mathbb{R})$.

Proposition 5 (Connections between expectation, covariance, and orthogonality).

a. X is orthogonal to itself, or any scalar multiple cX, iff it is equal to zero.

b. Zero-mean is equivalent to orthogonal to 1.

c. The inner product is equal to the covariance iff one or both of X and Y is zero-mean.

d. As a special case, the norm is equal to the standard deviation iff X is zero-mean.

e. $\cos\theta$ is equal to the correlation coefficient ρ when X and Y are both zero-mean.

f. Uncorrelated random variables X and Y are orthogonal iff X or Y is zero-mean.

As exercises in becoming familiar with the definitions, verify that $\mathcal{H}$ is actually a real inner product space, and prove Proposition 5 above.

The zero-mean assumption is so helpful that we may also consider the space of **centered** or *demeaned* random variables $\mathcal{H}/\mathbb{R}$, in which two random variables are equivalent if they differ by a constant. Then the picture is greatly simplified: covariance is the inner product, standard deviation is the norm, uncorrelatedness is orthogonality, and all affine functions are linear.

One final interesting connection: independence is often denoted $X \perp\!\!\!\perp Y$ to indicate an intuitive orthogonality in the information given by X and Y. Here, though, independence implies uncorrelatedness, which gives orthogonality $X \perp Y$ in the zero-mean case, and orthogonality then always implies *linear* independence.

■