An Introduction to Inner-Product Spaces of Random Variables

Henry D. Pfister

November 16, 2025

Contents

1 Introduction

In many areas of Electrical and Computer Engineering (ECE), including signal processing, communications, and control, we work with vector spaces equipped with additional structure. An *inner-product space* allows us to define angles, lengths, and orthogonality, extending the familiar geometry of $\mathbb{R}^n$ to more abstract settings.

2 Background: Fields, Vectors, and Probability

Definition 2.1 (Field)**.** A field $\mathbb{F}$ is a set equipped with two binary operations, addition $(+)$ and multiplication $(\cdot)$, satisfying the following axioms. There exist distinct elements $0, 1 \in \mathbb{F}$ such that:

(F1) $(\mathbb{F}, +)$ forms an abelian group: for all $a, b, c \in \mathbb{F}$, we have $a + (b + c) = (a + b) + c$, and $a + b = b + a$; there is an additive identity $0 \in \mathbb{F}$ such that $a + 0 = a$ and, for each $a \in \mathbb{F}$, there exists $-a \in \mathbb{F}$ with $a + (-a) = 0$.

(F2) $(\mathbb{F}\setminus\{0\}, \cdot)$ forms an abelian group: for all $a, b, c \in \mathbb{F}$ with $a, b, c \neq 0$, we have $a \cdot (b \cdot c) = (a \cdot b) \cdot c$, $a \cdot b = b \cdot a$; there is a multiplicative identity $1 \in \mathbb{F}$ such that $1 \cdot a = a$ and, for each $a \in \mathbb{F}\setminus\{0\}$, there exists $a^{-1} \in \mathbb{F}$ with $a \cdot a^{-1} = 1$.

(F3) Multiplication distributes over addition: for all $a, b, c \in \mathbb{F}$, $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$.

Example 2.1 (Field Examples)**.** The following standard examples satisfy the field axioms:

- The real numbers $\mathbb{R}$ with standard addition and multiplication.
- The complex numbers $\mathbb{C}$ with standard addition and multiplication.
- The set of integers $\{0, 1, \ldots, p-1\}$ for prime p with addition and multiplication modulo p.

Definition 2.2 (Vector Space)**.** A vector space V over a field $\mathbb{F}$ (e.g., $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) is a set equipped with two operations:

(i) Vector addition: $u + v \in V$ for all $u, v \in V$.

(ii) Scalar multiplication: $\alpha v \in V$ for all $\alpha \in \mathbb{F}$, $v \in V$.

These satisfy the following axioms for all $u, v, w \in V$ and $\alpha, \beta \in \mathbb{F}$:

(V1) Vectors with addition $(V, +)$ form an abelian group: for all $u, v, w \in V$, $u+(v+w) = (u+v)+w$, and $u + v = v + u$; there exists an additive identity vector $0 \in V$ such that $u + 0 = u$ and, for each $u \in V$, there exists $-u \in V$ with $u + (-u) = 0$.

(V2) $1 \cdot u = u$ (scalar identity)

(V3) $\alpha(\beta u) = (\alpha\beta)u$ (scalar associativity)

(V4) $(\alpha + \beta)u = \alpha u + \beta u$ (scalar distributivity)

(V5) $\alpha(u + v) = \alpha u + \alpha v$ (vector distributivity)

Example 2.2 (Vector Space Examples)**.** The following familiar spaces satisfy the vector space axioms:

- The Euclidean spaces $\mathbb{R}^n$ and $\mathbb{C}^n$: Closure and all axioms follow component wise from the field laws of $\mathbb{R}$ or $\mathbb{C}$. For instance, vector addition is defined by $(u + v)_i = u_i + v_i$ and scalar multiplication by $(\alpha u)_i = \alpha u_i$. So, the axioms reduce to the corresponding scalar identities.
- The set $\mathcal{F}$ of functions mapping a set $\mathcal{X}$ to $\mathbb{R}^n$ (or $\mathbb{C}^n$): For $f, g \in \mathcal{F}$ and $\alpha \in \mathbb{F}$, vector addition and scalar multiplication are defined by $(\alpha f + g)(x) = \alpha f(x) + g(x) \in \mathcal{F}$. Again, the axioms follow from the corresponding scalar identities.
- Real random variables $\mathcal{R}$: If $X, Y \in \mathcal{R}$ and $\alpha \in \mathbb{R}$, then $\alpha X + Y \in \mathcal{R}$ and the axioms follow from the corresponding scalar identities. Since random variables are formally defined as functions from $\Omega \to \mathbb{R}$, this follows from the previous item.
- Random vectors in $\mathbb{R}^n$: Similar to random variables, if X and Y are in this set and $\alpha \in \mathbb{R}$, then $\alpha X + Y$ is in this set. This also follows from the fact that functions from $\Omega \to \mathbb{R}^n$ form a vector space.

Definition 2.3 (Probability space)**.** A probability space is defined by a tuple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the sample space, a subset $E \subseteq \Omega$ is an event, $\mathcal{F}$ is a special collection of events, and $\mathbb{P} : \mathcal{F} \to [0, 1]$ is the probability function. The collection $\mathcal{F}$ of events, known as a σ-algebra, must satisfy $\Omega \in \mathcal{F}$, $A^c \in \mathcal{F}$ if $A \in \mathcal{F}$, and $\cup_{i \in I} A_i \in \mathcal{F}$ if $A_i \in \mathcal{F}$ for all $i \in I$. The probability function must satisfy:

(P1) **Nonnegativity:** $\mathbb{P}(A) \geq 0$.

(P2) **Normalization:** $\mathbb{P}(\Omega) = 1$.

(P3) **Additivity:** for disjoint A_i, $\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)$.

Definition 2.4 (Random variable)**.** A real random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is defined by a mapping $X\colon \Omega \to \mathbb{R}$ where $X(\omega)$ gives the value of X for the outcome $\omega \in \Omega$.

3 Inner Products, Norms, and Metrics

Definition 3.1 (Inner Product)**.** An inner product on V is a map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F}$ satisfying for all $u, v, w \in V$ and $\alpha \in \mathbb{F}$:

(IP1) $\langle u + w, v \rangle = \langle u, v \rangle + \langle w, v \rangle$ (linearity in first argument)

(IP2) $\langle \alpha u, v \rangle = \alpha \langle u, v \rangle$

(IP3) $\langle v, u \rangle = \overline{\langle u, v \rangle}$ (conjugate symmetry)

(IP4) $\langle v, v \rangle \geq 0$ with equality iff $v = 0$ (positive-definite)

For complex spaces, we use the above conjugate-symmetry convention. For random variables, we assume the field is $\mathbb{R}$ so conjugation is irrelevant.

Definition 3.2 (Norm)**.** A norm on V is a map $\| \cdot \| : V \to [0, \infty)$ satisfying for all $u, v \in V$ and $\alpha \in \mathbb{F}$:

(N1) $\|v\| \geq 0$ and $\|v\| = 0$ iff $v = 0$ (positive-definite)

(N2) $\|\alpha v\| = |\alpha| \, \|v\|$ (homogeneity)

(N3) $\|u + v\| \leq \|u\| + \|v\|$ (triangle inequality)

The standard induced distance metric between vectors is given by $d(u, v) = \|u - v\|$.

4 Cauchy–Schwarz and Induced Norms

Theorem 4.1 (Cauchy–Schwarz Inequality)**.** *For any $u, v \in V$, we have*

$$|\langle u, v \rangle|^2 \leq \langle u, u \rangle \, \langle v, v \rangle.$$

Proof. If $v = 0$, inequality holds trivially. Otherwise, for any scalar α, positivity gives

$$0 \leq \langle u - \alpha v, u - \alpha v \rangle = \langle u, u \rangle - \alpha \langle v, u \rangle - \overline{\alpha} \langle u, v \rangle + |\alpha|^2 \langle v, v \rangle.$$

Completing the square, we rewrite the right-hand side as

$$\langle v, v \rangle \left| \alpha - \frac{\langle v, u \rangle}{\langle v, v \rangle} \right|^2 \; + \; \langle u, u \rangle \; - \; \frac{|\langle u, v \rangle|^2}{\langle v, v \rangle}.$$

The first term is nonnegative and is zero exactly when $\alpha = \langle v, u \rangle / \langle v, v \rangle$, so this choice minimizes the expression. Substituting that value yields

$$0 \leq \langle u, u \rangle - \frac{|\langle u, v \rangle|^2}{\langle v, v \rangle},$$

which rearranges to $|\langle u, v \rangle|^2 \leq \langle u, u \rangle \, \langle v, v \rangle$. □

Theorem 4.2 (Induced Norm)**.** $\|v\| = \sqrt{\langle v, v\rangle}$ *satisfies all norm axioms.*

Proof. Positivity and homogeneity are immediate. Triangle inequality follows from Cauchy–Schwarz:

$$\|u+v\|^2 = \langle u+v, u+v\rangle = \|u\|^2 + \|v\|^2 + 2\Re\langle u, v\rangle \le \|u\|^2 + \|v\|^2 + 2\|u\|\|v\| = (\|u\| + \|v\|)^2. \quad \square$$

5 Examples of Inner-Product Spaces

Example 5.1 (Real Euclidean Space)**.** On $\mathbb{R}^n$, the standard inner product is $\langle x, y\rangle = x^\top y = \sum_i x_i y_i$, where $x^\top$ denotes the transpose of the column vector x. It is linear in the first argument, symmetric since $x^\top y = y^\top x$, and positive-definite because $x^\top x = \sum_i x_i^2 \ge 0$ with equality iff $x = 0$.

Example 5.2 (Complex Euclidean Space)**.** On $\mathbb{C}^n$, the standard inner product is $\langle x, y\rangle = y^* x = \sum_i \overline{y_i} x_i$, where y^* denotes the conjugate (or Hermitian) transpose of y. It is linear in x, conjugate symmetric in y (because $y^*x = \overline{x^* y}$), and positive-definite since $x^*x = \sum |x_i|^2 \ge 0$ with equality iff $x = 0$.

Example 5.3 (Weighted Complex Inner Product)**.** On $\mathbb{C}^n$, for W that is Hermitian and positive definite (i.e., $W = W^*$ and $x^*Wx > 0$ for $x \neq 0$), define $\langle x, y\rangle_W = y^*Wx$. This is linear in x, conjugate symmetric because $\langle y, x\rangle_W = x^*Wy = \overline{y^*Wx}$, and positive-definite because $x^*Wx > 0$.

Example 5.4 (Random Variables)**.** For the vector space of real random variables, define the inner product $\langle X, Y\rangle = \mathbb{E}[YX]$. By convention, the inner product space only includes vectors with finite induced norm (i.e., real random variables with finite second moment). Since X, Y are real random variables and the scalar field is $\mathbb{R}$, linearity in the both arguments follows from the linearity of expectation. In addition, symmetry follows from the commutativity of multiplication and positive-definiteness holds because $\mathbb{E}[X^2] \ge 0$ with equality iff $X = 0$ almost surely (note: probability only distinguishes random variables up to the condition of almost sure equality given by $\mathbb{P}(X \neq Y) = 0$).

Since real random variables form a vector space, the key condition to check is that the subspace with finite-second moment is closed (i.e., that, $\alpha X + Y$ has a finite second moment if X and Y have finite second moments). For this, we can use Cauchy-Schwarz to write

$$\mathbb{E}[(\alpha X + Y)^2] = \alpha^2\mathbb{E}[X^2] + 2\alpha\mathbb{E}[XY] + \mathbb{E}[Y^2] \le \alpha^2\mathbb{E}[X^2] + 2\alpha\sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]} + \mathbb{E}[Y^2] < \infty.$$

Example 5.5 (Random Vectors)**.** For real random vectors $\mathbf{X}, \mathbf{Y}$ with finite covariance matrices[1], define

$$\langle \mathbf{X}, \mathbf{Y}\rangle \;=\; \mathbb{E}[\mathbf{Y}^\top \mathbf{X}] \;=\; \sum_{i=1}^{n} \mathbb{E}[Y_i X_i].$$

Linearity in both arguments and symmetry follow from the real scalar field, linearity of expectation, and commutativity of multiplication. Positive-definiteness holds because

$$\langle \mathbf{X}, \mathbf{X}\rangle = \mathbb{E}[\mathbf{X}^\top \mathbf{X}] = \mathbb{E}\big[\|\mathbf{X}\|^2\big] \ge 0,$$

with equality iff $\mathbf{X} = 0$ almost surely; thus we identify random vectors that are equal almost surely. The induced norm equals the expected squared length

$$\|\mathbf{X}\|^2 = \mathbb{E}\big[\|\mathbf{X}\|^2\big].$$

Since the space of random vectors forms a vector space, one can similarly verify that the subspace with finite covariance matrices is closed using the Cauchy-Schwarz inequality.

[1] We denote the covariance of a real random vector by $\mathrm{Cov}(\mathbf{X}) \coloneqq \mathbb{E}\big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top\big]$.

Summary Table:

Space	Inner Product	Norm
$\mathbb{R}^n$	$x^\top y$	$\sqrt{x^\top x}$
$\mathbb{C}^n$	$x^* y$	$\sqrt{x^* x}$
Weighted	$x^* W y$	$\sqrt{x^* W x}$
Random Variables	$\mathbb{E}[YX]$	$\sqrt{\mathbb{E}[X^2]}$
Random Vectors	$\mathbb{E}[\mathbf{Y}^\top \mathbf{X}]$	$\sqrt{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]}$

6 Orthogonality, Projections, and Gram-Schmidt

6.1 Orthogonality and Projection

Definition 6.1 (Orthogonality and Orthogonal Complement)**.** In an inner-product space, vectors u and w are orthogonal, written $u \perp w$, if $\langle u, w\rangle = 0$. For a subspace $W \subseteq V$, the orthogonal complement is

$$W^\perp \;=\; \{\, v \in V :\; \langle v, w\rangle = 0 \text{ for all } w \in W \,\}.$$

Definition 6.2 (Orthonormal Set and Basis)**.** A collection $\{w_1, \dots, w_k\} \subseteq V$ is orthonormal if $\langle w_i, w_j\rangle = \delta_{ij}$. If additionally $\operatorname{span}\{w_1, \dots, w_k\} = W$, then it is an orthonormal basis of the subspace W.

Definition 6.3 (Standard Basis)**.** For $V = \mathbb{R}^n$ and $V = \mathbb{C}^n$, the standard orthonormal basis is given by the set of vectors, $\{e_1, e_2, \dots, e_n\}$, where e_i is all zero except for 1 in the i-th position.

Definition 6.4 (Orthogonal Projection)**.** Let $W \subseteq V$ be a subspace. The (orthogonal) projection of $v \in V$ onto W is defined by

$$P_W(v) \in \arg\min_{w \in W} \|v - w\|.$$

If V is finite dimensional, the minimum is achieved uniquely and called the *best approximation* of v by vectors in W. Equivalently, $r = v - P_W(v)$ is characterized by $r \perp W$.

Theorem 6.1 (Projection Theorem)**.** *Let V be a finite-dimensional inner-product space and $W \subseteq V$ a subspace. For every $v \in V$ there exists a unique decomposition*

$$v \;=\; w + r, \qquad w \in W,\; r \in W^\perp.$$

Thus, $P_W(v) = w$ is the unique minimizer of $\|v - w\|$ over W. If $\{w_i\}_{i=1}^k$ is an orthonormal basis of W, then

$$P_W(v) \;=\; \sum_{i=1}^{k} \langle v, w_i\rangle\, w_i, \qquad \|v - P_W(v)\|^2 \;=\; \|v\|^2 - \sum_{i=1}^{k} |\langle v, w_i\rangle|^2.$$

Proof. Let $w = \sum_{i=1}^k \langle v, w_i\rangle\, w_i$ and set $r = v - w$. For each j,

$$\langle r, w_j\rangle = \langle v, w_j\rangle - \sum_{i=1}^{k} \langle v, w_i\rangle\langle w_i, w_j\rangle = \langle v, w_j\rangle - \langle v, w_j\rangle = 0,$$

so $r \in W^{\perp}$, yielding a decomposition $v = w + r$. For any $\tilde{w} = \sum_i c_i w_i \in W$,

$$\|v - \tilde{w}\|^2 = \Big\| r + \sum_i (\langle v, w_i \rangle - c_i) w_i \Big\|^2 = \|r\|^2 + \sum_i |c_i - \langle v, w_i \rangle|^2 \;\geq\; \|r\|^2,$$

with equality iff $c_i = \langle v, w_i \rangle$ for all i, i.e., $\tilde{w} = w$. Hence $w = P_W(v)$ is the unique minimizer. The length of the sum of orthogonal vectors is given by the Pythagorean theorem: $\|v\|^2 = \|w\|^2 + \|r\|^2$ and $\|w\|^2 = \sum_i |\langle v, w_i \rangle|^2$. □

Corollary 6.1 (Projection onto nested subspaces)**.** *Let V be a finite-dimensional inner-product space and let $U \subseteq W \subseteq V$ be subspaces. Then, for every $v \in V$, we have*

$$P_U\big(P_W(v)\big) \;=\; P_U(v).$$

Proof. By Theorem 6.1, write $v = w + r$ with $w \in W$ and $r \in W^{\perp}$. Because $U \subseteq W$, we have $r \perp U$, so $P_U(r) = 0$. Hence, we find that

$$P_U(v) \;=\; P_U(w + r) \;=\; P_U(w) \;=\; P_U\big(P_W(v)\big).$$ □

6.2 Gram–Schmidt Orthogonalization

Given linearly independent vectors $v_1, \ldots, v_k$ in an inner-product space V, define

$$u_1 = v_1, \qquad w_1 = \frac{u_1}{\|u_1\|},$$

and for $j = 2, \ldots, k$ set

$$u_j \;=\; v_j - \sum_{i=1}^{j-1} \langle v_j, w_i \rangle \, w_i, \qquad w_j \;=\; \frac{u_j}{\|u_j\|} \;\text{ whenever } u_j \neq 0.$$

If some $u_j = 0$ for $j \leq k$, then $v_j \in \operatorname{span}\{v_1, \ldots, v_{j-1}\}$ and the set is not linearly independent (i.e., the assumed starting condition has been violated).

To handle sets of possibly linearly dependent vectors, one can instead move linearly dependent vectors to the end of the list and renumber. When all linearly independent vectors have been processed, one has found an orthonormal spanning set. To complete the process, one can simply project the remaining vectors onto this orthonormal set.

Theorem 6.2 (Properties of Gram–Schmidt)**.** *If $v_1, \ldots, v_k$ are linearly independent, then $\{w_1, \ldots, w_k\}$ is orthonormal and $\operatorname{span}\{w_1, \ldots, w_k\} = \operatorname{span}\{v_1, \ldots, v_k\}$. For each j, $u_j \perp \operatorname{span}\{w_1, \ldots, w_{j-1}\}$ and*

$$v_j \;=\; \sum_{i=1}^{j} \langle v_j, w_i \rangle \, w_i.$$

Proof sketch. By construction, u_j is obtained by subtracting from v_j its projection onto $\operatorname{span}\{w_1, \ldots, w_{j-1}\}$, so $u_j \perp w_i$ for $i < j$. Normalizing gives $\langle w_i, w_j \rangle = \delta_{ij}$. Inductively, we have $\operatorname{span}\{w_1, \ldots, w_j\} = \operatorname{span}\{v_1, \ldots, v_j\}$, which yields the stated decomposition of v_j. □

Example 6.1 (Projection in $\mathbb{R}^3$)**.** Let

$$W = \operatorname{span}\{a, b\}, \quad a = (1, 1, 0)^{\top}, \; b = (1, 0, 0)^{\top}.$$

First, apply Gram–Schmidt to obtain an orthonormal basis for W:

$$w_1 = \frac{a}{\|a\|} = \frac{1}{\sqrt{2}}(1,1,0)^\top, \quad u_2 = b - \langle b, w_1\rangle w_1 = b - \frac{1}{\sqrt{2}}\, w_1 = (\tfrac{1}{2}, -\tfrac{1}{2}, 0)^\top,$$

$$w_2 = \frac{u_2}{\|u_2\|} = \frac{(\frac{1}{2}, -\frac{1}{2}, 0)}{\sqrt{\frac{1}{2}}} = \Big(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0\Big)^\top.$$

Given $v = (2, -1, 3)^\top$, the projection onto W is

$$P_W(v) = \langle v, w_1\rangle w_1 + \langle v, w_2\rangle w_2.$$

Next, compute the coefficients:

$$\langle v, w_1\rangle = \frac{1}{\sqrt{2}}(2 + (-1) + 0) = \frac{1}{\sqrt{2}}, \qquad \langle v, w_2\rangle = \frac{\sqrt{2}}{2}(2) + \Big(-\frac{\sqrt{2}}{2}\Big)(-1) = \frac{3\sqrt{2}}{2}.$$

This gives

$$P_W(v) = \frac{1}{\sqrt{2}}\,\frac{1}{\sqrt{2}}(1,1,0) + \frac{3\sqrt{2}}{2}\Big(\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0\Big) = \Big(\frac{1}{2}, \frac{1}{2}, 0\Big) + \Big(\frac{3}{2}, -\frac{3}{2}, 0\Big) = (2, -1, 0)^T$$

and the residual $r = v - P_W(v) = (0, 0, 3)^T$ is orthogonal to W since $\langle r, w_i\rangle = 0$ for $i = 1, 2$.

6.3 Normal Equations

Suppose V is an inner-product space and the subspace W is spanned by $w_1, \ldots, w_k \in V$. Consider the situation where the sequence $w_1, \ldots, w_k$ is linearly independent, but not orthogonal. In this case, it is not possible to apply Theorem 6.1 directly. It is nevertheless possible to obtain a similar expression for the best approximation of v by vectors in W. Theorem 6.1 shows that $w \in W$ is a best approximation of $v \in V$ by vectors in W if and only if $v - w$ is orthogonal to every vector in W. This implies that

$$\langle v - w,\, w_j\rangle = \left\langle v - \sum_{i=1}^{k} s_i w_i,\, w_j \right\rangle = 0$$

or, equivalently,

$$\sum_{i=1}^{k} s_i \,\langle w_i,\, w_j\rangle = \langle v,\, w_j\rangle$$

for $j = 1, \ldots, k$. These conditions yield a system of k linear equations in k unknowns, which can be written in the matrix form

$$\underbrace{\begin{bmatrix} \langle w_1, w_1\rangle & \langle w_2, w_1\rangle & \cdots & \langle w_k, w_1\rangle \\ \langle w_1, w_2\rangle & \langle w_2, w_2\rangle & \cdots & \langle w_k, w_2\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle w_1, w_k\rangle & \langle w_2, w_k\rangle & \cdots & \langle w_k, w_k\rangle \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{bmatrix}}_{s} = \underbrace{\begin{bmatrix} \langle v, w_1\rangle \\ \langle v, w_2\rangle \\ \vdots \\ \langle v, w_k\rangle \end{bmatrix}}_{t}.$$

We can rewrite this matrix equation as

$$Gs = t,$$

where G is called a *Gramian* matrix, t is called a *cross-correlation vector*, and $s^T = (s_1, s_2, \ldots, s_k)$ is the vector of coefficients. Equations of this form are collectively known as *normal equations*.

Definition 6.5. A matrix $M \in \mathbb{C}^{k\times k}$ is *positive-semidefinite* if $M^* = M$ and $v^*Mv \geq 0$ for all $v \in \mathbb{C}^k - \{0\}$. If the inequality is strict, M is *positive-definite*.

An important aspect of positive-definite matrices is that they are always invertible. This follows from noting that $Mv = 0$ for $v \neq 0$ implies that $v^*Mv = 0$ and contradicts the definition of positive definite.

Theorem 6.3. *A Gramian matrix G is always positive-semidefinite. It is positive-definite if and only if the vectors $w_1, \ldots, w_k$ are linearly independent.*

Proof. Since $g_{ij} = \langle w_j, w_i \rangle$, the conjugation property of the inner product implies $G^* = G$. For any $v = (v_1, \ldots, v_k) \in \mathbb{C}^k$, we can write

$$\begin{aligned} v^*Gv &= \sum_{i=1}^{k}\sum_{j=1}^{k} \bar{v}_i g_{ij} v_j = \sum_{i=1}^{k}\sum_{j=1}^{k} \bar{v}_i \langle w_j, w_i \rangle v_j \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} \langle v_j w_j, v_i w_i \rangle = \left\langle \sum_{j=1}^{k} v_j w_j, \sum_{i=1}^{k} v_i w_i \right\rangle \\ &= \left\| \sum_{i=1}^{k} v_i w_i \right\|^2 \geq 0, \end{aligned} \tag{1}$$

with equality if and only if $\sum_{i=1}^{k} v_i w_i = 0$. Since v is arbitrary, equality may occur iff $v = 0$ or $w_1, \ldots, w_k$ are linearly dependent. □

Proposition 6.1 (Normal equations for projection onto column space)**.** *Consider $A \in \mathbb{R}^{n\times k}$ with column space $W = \text{range}(A)$ and let $b \in \mathbb{R}^n$. Then, the orthogonal projection of b onto W is $P_W(b) = A\hat{x}$ where $\hat{x}$ satisfies the normal equations*

$$A^\top A\, \hat{x} \;=\; A^\top b.$$

If $n \geq k$ and A has full rank, then solution is unique and satisfies $\hat{x} = (A^\top A)^{-1} A^\top b$.

Proof. By the Projection Theorem, $r = b - A\hat{x}$ is orthogonal to W, so $A^\top r = 0$. Thus, $A^\top(b - A\hat{x}) = 0$, which yields the stated form of the normal equations. If $n \geq k$ and A is full rank, then $Ax = 0$ iff $x = 0$. Thus, $\|Ax\|^2 = (Ax)^\top Ax = x^\top (A^\top A)x \geq 0$ with equality iff $x = 0$. This implies that $A^\top A$ is positive definite and hence invertible. □

Remark 6.1 (QR decomposition via Gram–Schmidt)**.** For the standard spaces $V = \mathbb{R}^n$ and $V = \mathbb{C}^n$, we can define the matrix $A = [\, v_1 \;\cdots\; v_k \,] \in \mathbb{F}^{n\times k}$, where each column is an input vector. Applying Gram–Schmidt to this set of vectors naturally produces $Q = [\, w_1 \;\cdots\; w_k \,]$ with $Q^*Q = I$ and an upper-triangular $R \in \mathbb{F}^{k\times k}$ with entries $r_{ij} = \langle v_j, w_i \rangle$ for $i \leq j$ (and $r_{ij} = 0$ for $i > j$). This is the classical QR factorization of the matrix A. From a numerical perspective, one can also reorder the computations to get what is known as the modified Gram–Schmidt procedure. This is algebraically equivalent but can reduce the loss in precision due to round-off error.

7 Projection and Orthogonality for Random Variables

The procedures described above can be applied directly to inner-product spaces of random variables. In this section, we provide probabilistic interpretations of these operations.

7.1 Linear Combinations of Random Variables

Proposition 7.1 (Orthonormal sets are uncorrelated with unit energy)**.** *Let $X_1, \dots, X_k$ be random variables in the standard inner-product space of real random variables. We can apply Gram–Schmidt to obtain the orthonormal set $Y_i = \sum_{j=1}^{k} a_{ij} X_j$. Then, by orthonormality, for all i, j,*

$$\mathbb{E}[Y_i Y_j] \;=\; \langle Y_i, Y_j \rangle \;=\; \delta_{ij}.$$

Thus, the Y_i are pairwise uncorrelated and unit energy. In vector form, $\mathbf{Y} = A\mathbf{X}$ for an invertible upper-triangular matrix A determined by the procedure, so the orthonormal set of variables is obtained by a linear transform of the original set.

Proposition 7.2 (The constant random variable)**.** *If we prepend the constant random variable* 1 *to the list, i.e., start with $X_0 = 1$, then $Y_0 = \frac{1}{\|1\|} = \frac{1}{\sqrt{\mathbb{E}[1]}} = 1$. For any subsequent variable X, the first step in Gram-Schmidt outputs*

$$Z \;=\; X - \langle X, Y_0 \rangle Y_0 \;=\; X - \mathbb{E}[X],$$

so Z has zero mean and is orthogonal to Y_0. Continuing Gram–Schmidt with these centered variables yields an orthonormal set $1, Y_1, Y_2, \dots$ in which every Y_i for $i \geq 1$ has mean zero and the Y_i are pairwise uncorrelated.

Example 7.1. Consider the space of real random variables with finite second moment and suppose that $\mathbb{E}[X_1] = 2$, $\mathbb{E}[X_2] = -1$, $\mathrm{Var}(X_1) = 4$, $\mathrm{Var}(X_2) = 9$, and $\mathrm{Cov}(X_1, X_2) = 3$. We will include the constant 1 and perform Gram–Schmidt as follows:

$$Y_0 = 1, \qquad Y_1 = \frac{X_1 - \mathbb{E}[X_1]}{\sqrt{\mathrm{Var}(X_1)}} = \frac{X_1 - 2}{2}.$$

Next, project X_2 onto $\mathrm{span}\{Y_0, Y_1\}$ and subtract:

$$\langle X_2, Y_0 \rangle = \mathbb{E}[X_2] = -1, \quad \langle X_2, Y_1 \rangle = \mathbb{E}\left[X_2 \, \frac{X_1 - 2}{2} \right] = \frac{\mathrm{Cov}(X_1, X_2)}{2} = \frac{3}{2},$$

$$Y_2' \;=\; X_2 - \langle X_2, Y_0 \rangle Y_0 - \langle X_2, Y_1 \rangle Y_1 \;=\; X_2 + 1 - \frac{3}{2} \cdot \frac{X_1 - 2}{2} \;=\; X_2 - \frac{3}{4} X_1 + \frac{5}{2}.$$

The random variable Y_2' has zero mean and is orthogonal to Y_1 by construction. Thus, its norm is

$$\|Y_2'\|^2 \;=\; \mathrm{Var}\big(X_2 - \tfrac{3}{4} X_1\big) \;=\; \mathrm{Var}(X_2) + \left(\tfrac{3}{4}\right)^2 \mathrm{Var}(X_1) - 2 \cdot \tfrac{3}{4} \, \mathrm{Cov}(X_1, X_2) \;=\; 9 + \tfrac{9}{4} - \tfrac{9}{2} \;=\; \tfrac{27}{4}.$$

Hence, we find that

$$Y_2 \;=\; \frac{Y_2'}{\|Y_2'\|} \;=\; \frac{2}{3\sqrt{3}} \left(X_2 - \frac{3}{4} X_1 + \frac{5}{2} \right).$$

From $\mathbb{E}[Y_1] = \mathbb{E}[Y_2] = 0$ and $\mathbb{E}[Y_1 Y_2] = 0$, it follows that $\{1, Y_1, Y_2\}$ is an orthonormal set and Y_1, Y_2 are uncorrelated zero-mean random variables.

In the inner-product space of real random variables, the Gramian matrix for a set of random variables $X_1, \dots, X_k$ is given by

$$G = \begin{bmatrix} \langle X_1,\, X_1 \rangle & \langle X_2,\, X_1 \rangle & \cdots & \langle X_k,\, X_1 \rangle \\ \langle X_1,\, X_2 \rangle & \langle X_2,\, X_2 \rangle & \cdots & \langle X_k,\, X_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle X_1,\, X_k \rangle & \langle X_2,\, X_k \rangle & \cdots & \langle X_k,\, X_k \rangle \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1 X_1] & \mathbb{E}[X_1 X_2] & \cdots & \mathbb{E}[X_1 X_k] \\ \mathbb{E}[X_2 X_1] & \mathbb{E}[X_2 X_2] & \cdots & \mathbb{E}[X_2 X_k] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[X_k X_1] & \mathbb{E}[X_k X_2] & \cdots & \mathbb{E}[X_k X_k] \end{bmatrix}.$$

Since $\mathbb{E}[X_i X_j] = \mathrm{Cov}(X_i, X_j) + \mathbb{E}[X_i]\mathbb{E}[X_j]$, we see that $G = \mathrm{Cov}(\mathbf{X}) + \mu\mu^\top$.

Proposition 7.3 (Normal equations for best affine approximation of a random variable)**.** *Let* $\mathbf{X} = (X_1, \ldots, X_k)$ *and* Y *be real random variables with finite second moments. We define*

$$\begin{aligned}
&\mathbf{X} = (X_1, \ldots, X_k)^\top, \\
&\mu = (\mu_1, \ldots, \mu_k)^\top = \mathbb{E}[\mathbf{X}], && \mu_i = \mathbb{E}[X_i], \\
&\Sigma = \mathrm{Cov}(\mathbf{X}) = \mathbb{E}\big[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^\top\big], && \Sigma_{i,j} = \mathrm{Cov}(X_i, X_j), \\
&p = (p_1, \ldots, p_k)^\top = \mathbb{E}\big[(\mathbf{X} - \mu)\,(Y - \mathbb{E}[Y])\big], && p_i = \mathrm{Cov}(X_i, Y).
\end{aligned}$$

Then, the orthogonal projection of Y *onto* $\mathrm{span}\{1, X_1, \ldots, X_k\}$ *has the form*

$$\widehat{Y} \;=\; c + w^\top \mathbf{X},$$

where vectors w *and* c *satisfy the normal equations*

$$\Sigma\, w \;=\; p, \qquad c \;=\; \mathbb{E}[Y] - w^\top \mu.$$

If Σ *is invertible, then* $w = \Sigma^{-1} p$*. Otherwise, all solutions are optimal but the minimum-norm solution is typically preferred.*

Proof sketch. Let $R = Y - c - w^\top \mathbf{X}$ be the residual random variable. Orthogonality of R to each basis element $1, X_1, \ldots, X_k$ yields $\mathbb{E}[R] = 0$ and $\mathbb{E}[R(\mathbf{X} - \mu)] = 0$. After removing the means, the augmented form is the usual least-squares normal equations $\Sigma w = p$. Thus, the solution follow from solving the normal equations for the centered random variables. □

Example 7.2 (Approximation of e^X for $X \sim \mathcal{N}(0,1)$)**.** Find the best mean-square approximation of $Y = e^X$ by $\mathrm{span}\{1, X, X^2\}$ for $X \sim \mathcal{N}(0,1)$:

$$\widehat{Y} \;=\; a_0 + a_1 X + a_2 X^2.$$

The normal equations enforce orthogonality of the residual to $1, X, X^2$:

$$\mathbb{E}[Y] = a_0 + a_2\,\mathbb{E}[X^2], \quad \mathbb{E}[XY] = a_1\,\mathbb{E}[X^2] + a_2\,\mathbb{E}[X^3], \quad \mathbb{E}[X^2 Y] = a_0\,\mathbb{E}[X^2] + a_2\,\mathbb{E}[X^4].$$

Using $\mathbb{E}[X] = 0$, $\mathbb{E}[X^2] = 1$, $\mathbb{E}[X^3] = 0$, $\mathbb{E}[X^4] = 3$ and the MGF $M_X(t) = \mathbb{E}[e^{tX}] = \exp(t^2/2)$,

$$\mathbb{E}[Y] = e^{1/2}, \quad \mathbb{E}[XY] = \frac{d}{dt} M_X(t)\Big|_{t=1} = e^{1/2}, \quad \mathbb{E}[X^2 Y] = \frac{d^2}{dt^2} M_X(t)\Big|_{t=1} = 2e^{1/2}.$$

Hence,

$$e^{1/2} = a_0 + a_2, \qquad e^{1/2} = a_1, \qquad 2e^{1/2} = a_0 + 3a_2,$$

which yields

$$a_0 = \tfrac{1}{2} e^{1/2}, \quad a_1 = e^{1/2}, \quad a_2 = \tfrac{1}{2} e^{1/2}.$$

Therefore,

$$\widehat{Y} \;=\; \tfrac{e^{1/2}}{2}\left(1 + 2X + X^2\right) \;=\; e^{1/2}\Big(\tfrac{1}{2} + X + \tfrac{1}{2} X^2\Big),$$

and $Y - \widehat{Y}$ is orthogonal to 1, X, and X^2, so $\widehat{Y}$ is the best quadratic mean-square approximation to e^X.

7.2 Functions of Random Variables

Now, we introduce the conditional expectation and the law of nested conditional expectation. These play a key role in the estimation of one random variable by functions of another random variable.

Definition 7.1 (Indicator random variable)**.** For a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an event $A \in \mathcal{F}$, the indicator random variable $\mathbf{1}_A$ is defined by

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

It follows that $\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{A\cap B}$, $\mathbb{E}[\mathbf{1}_A] = \mathbb{P}(A)$, $\mathbb{E}[\mathbf{1}_A \mathbf{1}_B] = \mathbb{P}(A \cap B)$, and $\mathbb{E}[X\, \mathbf{1}_A] = \mathbb{E}[X \mid A]\, \mathbb{P}(A)$.

Definition 7.2 (Standard basis)**.** For a probability space with countably many outcomes, the standard orthonormal basis is given by the set of normalized indicator random variables $\{E_\alpha\}_{\alpha\in\Omega}$, where

$$E_\alpha = \frac{\mathbf{1}_{\{\alpha\}}}{\sqrt{\mathbb{P}(\{\alpha\})}}.$$

They are orthonormal because

$$\mathbb{E}\left[E_\alpha E_\beta\right] = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{otherwise.} \end{cases}$$

Thus, any random variable X can be written uniquely as

$$X = \sum_{\omega\in\Omega} X(\omega)\mathbf{1}_{\{\omega\}} = \sum_{\omega\in\Omega} \left(\sqrt{\mathbb{P}(\omega)}X(\omega)\right) E_\omega.$$

Definition 7.3 (Conditional expectation)**.** Let X and Y be real random variables with $\mathbb{E}[X^2] < \infty$. Recall that, when Y is discrete and $\mathbb{P}(Y = y) > 0$, we can define

$$g(y) \;=\; \mathbb{E}[X \mid Y = y] \;=\; \sum_{x\in S_X} x\, \mathbb{P}(X = x \mid Y = y).$$

Similarly, when (X, Y) admits a joint density $f_{X,Y}$ and $f_Y(y) > 0$, we can define

$$g(y) \;=\; \mathbb{E}[X \mid Y = y] \;=\; \int_{-\infty}^{\infty} x\, f_{X|Y}(x \mid y)\, dx \;=\; \frac{\int x\, f_{X,Y}(x, y)\, dx}{f_Y(y)}.$$

The conditional expectation of X given Y is the random variable g(Y) and is denoted by $\mathbb{E}[X \mid Y]$.

Theorem 7.1 (Law of nested conditional expectation)**.** *For real random variables X, Y, Z with $\mathbb{E}[X^2] < \infty$, we have*

$$\begin{aligned} \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] &= \mathbb{E}[X], \\ \mathbb{E}\big[\mathbb{E}[X \mid Y, Z]\,\big|\, Y\big] &= \mathbb{E}[X \mid Y]. \end{aligned}$$

Proof. We only prove the case of discrete random variables. For the first expression, we see that

$$\begin{aligned} \mathbb{E}\big[\mathbb{E}[X \mid Y]\big] &= \sum_{y\in S_Y} \mathbb{E}[X \mid Y = y]\, \mathbb{P}(Y = y) \\ &= \sum_{y\in S_Y} \sum_{x\in S_X} x\, \mathbb{P}(X = x \mid Y = y)\, \mathbb{P}(Y = y) \\ &= \sum_{x\in S_X} x\mathbb{P}(X = x) \\ &= \mathbb{E}[X]. \end{aligned}$$

For the second expression, consider any y with $\mathbb{P}(Y = y) > 0$ and observe that

$$\begin{aligned}
\mathbb{E}\big[\mathbb{E}[X \mid Y, Z] \,\big|\, Y = y\big] &= \sum_{z \in S_Z} \mathbb{E}[X \mid Y = y, Z = z]\, \mathbb{P}(Z = z \mid Y = y) \\
&= \sum_{z \in S_Z} \sum_{x \in S_X} x\, \mathbb{P}(X = x \mid Y = y, Z = z)\, \mathbb{P}(Z = z \mid Y = y) \\
&= \sum_{x \in S_X} x \sum_{z \in S_Z} \mathbb{P}(X = x, Z = z \mid Y = y) \\
&= \sum_{x \in S_X} x\, \mathbb{P}(X = x \mid Y = y) \\
&= \mathbb{E}[X \mid Y = y].
\end{aligned}$$

Since this holds for all y with $\mathbb{P}(Y = y) > 0$, the implied random variables are also equivalent. $\square$

Definition 7.4 (Conditional variance)**.** Let X and Y be real random variables with $\mathbb{E}[X^2] < \infty$. The variance of X conditioned on an event A is the deterministic quantity denoted by

$$\operatorname{Var}(X \mid A) = \mathbb{E}\big[(X - \mathbb{E}[X \mid A])^2 \mid A\big] = \mathbb{E}[X^2 \mid A] - \mathbb{E}[X \mid A]^2.$$

Choosing the event $A = \{Y = y\}$, we define

$$h(y) = \operatorname{Var}(X \mid Y = y) = \mathbb{E}[X^2 \mid Y = y] - \mathbb{E}[X \mid Y = y]^2.$$

The conditional variance of X given Y is the random variable $h(Y)$ also defined by

$$\operatorname{Var}(X \mid Y) \coloneqq \mathbb{E}[X^2 \mid Y] - \mathbb{E}[X \mid Y]^2.$$

Definition 7.5 (Minimum mean-squared error)**.** Given real random variables X and Y with $\operatorname{Var}(X) < \infty$, an estimator of X given Y is a function $g\colon \mathbb{R} \to \mathbb{R}$. The mean-squared error (MSE) of the estimator g is defined to be

$$\operatorname{MSE}(g) \;=\; \mathbb{E}\big[(X - g(Y))^2\big].$$

The minimum mean-squared error (MMSE) over all estimators is

$$\operatorname{mmse}(X \mid Y) \;=\; \inf_{g\colon \mathbb{R} \to \mathbb{R}} \operatorname{MSE}(g),$$

where the infimum is over all functions $g\colon \mathbb{R} \to \mathbb{R}$. Any g achieving the minimum is called an MMSE estimator of X given Y.

Theorem 7.2 (MMSE estimate is conditional expectation)**.** *Given real random variables X and Y with $\operatorname{Var}(X) < \infty$, an MMSE estimator for X given Y exists and is given by*

$$g(y) \;=\; \mathbb{E}[X \mid Y = y],$$

Moreover, the minimum value satisfies

$$\operatorname{mmse}(X \mid Y) \;=\; \mathbb{E}\big[(X - \mathbb{E}[X \mid Y])^2\big] \;=\; \mathbb{E}\big[\operatorname{Var}(X \mid Y)\big].$$

Proof. We only prove the case of discrete random variables. For any y with $\mathbb{P}(Y = y) > 0$ and any estimator g, we have

$$\begin{aligned}\mathbb{E}\big[(X - g(Y))^2\big] &= \sum_{y \in S_Y} \mathbb{E}\big[(X - g(Y))^2 \mid Y = y\big]\, \mathbb{P}(Y = y) \\ &= \sum_{y \in S_Y} \mathbb{E}\big[\, X^2 - 2Xg(y) + g(y)^2 \mid Y = y\,\big]\, \mathbb{P}(Y = y).\end{aligned}$$

One can optimize separately over $g(y)$ for each $y \in S_Y$ and this gives the condition

$$0 = \frac{d}{d\, g(y)} \mathbb{E}\big[\, X^2 - 2Xg(y) + g(y)^2 \mid Y = y\big] = -2\mathbb{E}[X \mid Y = y] + 2g(y).$$

Since the stationary condition uniquely defines $g(y)$ and the second derivative is positive, the minimizer is given by

$$g(y) = \mathbb{E}[X \mid Y = y].$$

It follows that the MMSE is given by stated expression. □

Theorem 7.3. *In the inner-product space of variables with finite second moment, $\mathbb{E}[X \mid Y]$ is the orthogonal projection of X onto the subspace of random variables that are functions of Y.*

Example 7.3 (Conditional Expectation as Projection)**.** Let X, Y, Z be real random variables taking values in $\{0, 1, 2\}$ with $\mathbb{P}(X = x, Y = y) = 1/9$ for all pairs and, for each (x, y),

$$\mathbb{P}(Z = z \mid X = x, Y = y) = \begin{cases} 0.5, & z = x, \\ 0.3, & z = y, \\ 0.2, & z = 2, \\ 0, & \text{otherwise.} \end{cases}$$

Then, we have

$$\mathbb{E}[Z \mid X = x, Y = y] = 0.5\, x + 0.3\, y + 0.4,$$

and this need not lie in $\{0, 1, 2\}$; for example, $\mathbb{E}[Z \mid X = 2, Y = 1] = 1.7$.

Let V be the space of real random variables with inner product $\langle X, Y \rangle = \mathbb{E}[XY]$ for all $X, Y \in V$. One can view $W = \{g(X, Y) : g{:}\{0, 1, 2\}^2 \to \mathbb{R}\} \subseteq V$ as the subspace of V containing all random variables that are deterministic functions of X, Y. An orthonormal basis for W is given by

$$E_{a,b} = \frac{\mathbf{1}_{\{X=a, Y=b\}}}{\sqrt{\mathbb{P}(X = a, Y = b)}} = \sqrt{9}\, \mathbf{1}_{\{X=a, Y=b\}}, \qquad (a, b) \in \{0, 1, 2\}^2,$$

since the indicators are disjoint. The orthogonal projection of Z onto W is

$$\begin{aligned}
P_W(Z) &= \sum_{a,b} \langle Z, E_{a,b} \rangle \, E_{a,b} \\
&= \sum_{a,b} \mathbb{E}[Z \, E_{a,b}] \, E_{a,b} \\
&= \sum_{a,b} \frac{1}{\mathbb{P}(X=a, Y=b)} \mathbb{E}[Z \, \mathbf{1}_{\{X=a,Y=b\}}] \, \mathbf{1}_{\{X=a,Y=b\}} \\
&= \sum_{a,b} \frac{1}{\mathbb{P}(X=a, Y=b)} \big(\mathbb{E}[Z \,|\, X=a, Y=b] \mathbb{P}(X=a, Y=b)\big) \, \mathbf{1}_{\{X=a,Y=b\}} \\
&= \sum_{a,b} \mathbb{E}[Z \mid X=a, Y=b] \, \mathbf{1}_{\{X=a,Y=b\}} \\
&= \mathbb{E}[Z \mid X, Y].
\end{aligned}$$

By definition, this orthogonal projection is the element of W closest to Z in the sense that

$$\mathbb{E}[(Z - g(X,Y))^2] \geq \mathbb{E}[(Z - \mathbb{E}[Z \mid X, Y])^2],$$

for all $g \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Thus, $\mathbb{E}[Z \mid X, Y]$ is the best mean-square approximation of Z among all functions of (X, Y).

For random variables taking uncountably many different values, same result holds but more advanced math is needed to make the derivation rigorous. The standard approach is introduce measure theory and associate random variables with measurable functions. While this additional effort is useful for proofs, it does not provide much help in practice. One alternative in engineering is to generate an orthonormal basis for the space of polynomial functions (e.g., of X, Y) by applying Gram-Schmidt to a countable spanning set of polynomials (e.g., $1, X, Y, XY, X^2, Y^2, \ldots$). Since such polynomials span the space of L^2 functions, this is formally correct and practically useful (see Example 7.2).

Proposition 7.4 (Nested conditional expectations and projections)**.** *Consider the standard inner-product space of real random variables with elements X, Y, Z and the subspaces*

$$U \;=\; \{\, g(Y) : \, g \colon \mathbb{R} \to \mathbb{R} \,\}, \qquad W \;=\; \{\, h(Y,Z) : \, h \colon \mathbb{R}^2 \to \mathbb{R} \,\}.$$

Then, $U \subseteq W$ and orthogonal projections onto U and W define the conditional expectations:

$$P_U(X) = \mathbb{E}[X \mid Y], \qquad P_W(X) = \mathbb{E}[X \mid Y, Z].$$

By Corollary 6.1, projecting onto nested subspaces gives $P_U(P_W(X)) = P_U(X)$. In terms of random variables this implies that

$$\mathbb{E}\big[\, \mathbb{E}[X \mid Y, Z] \,\big|\, Y \big] \;=\; \mathbb{E}[X \mid Y].$$

For the case where Y is almost surely constant, we have $U = \operatorname{span}\{1\}$ and $\mathbb{E}[\mathbb{E}[X \mid Z]] = \mathbb{E}[X]$.

Theorem 7.4 (Law of total variance)**.** *Let X and Y be real random variables with $\mathbb{E}[X^2] < \infty$. Then, we have*

$$\operatorname{Var}(X) \;=\; \mathbb{E}\big[\operatorname{Var}(X \mid Y)\big] \;+\; \operatorname{Var}\big(\mathbb{E}[X \mid Y]\big).$$

Proof. For x in a Euclidean space with subspace W, the Pythagorean Theorem naturally implies $\|x\|^2 = \|x - P_W(x)\|^2 + \|P_W(x)\|^2$ because $x - P_W(x)$ and $P_W(x)$ are orthogonal. If we let W be the space of random variables determined by Y, then the law of total variance is given by projecting the random variable X onto W (which gives $P_W(X) = \mathbb{E}[X \,|\, Y]$) in the standard inner-product space of random variables. Using the properties of nested conditional expectation, this can be seen explicitly with

$$\begin{aligned}
\operatorname{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}\big[\mathbb{E}[X^2 \mid Y]\big] - \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]^2 \\
&= \mathbb{E}\big[\operatorname{Var}(X \mid Y) + \mathbb{E}[X \mid Y]^2\big] - \mathbb{E}\big[\mathbb{E}[X \mid Y]\big]^2 \\
&= \mathbb{E}\big[\operatorname{Var}(X \mid Y)\big] + \operatorname{Var}\big(\mathbb{E}[X \mid Y]\big).
\end{aligned}$$

□

Conditional expectation and Gram-Schmidt. There is subtle and often confusing relationship between Gram-Schmidt in the space of random variables and conditional expectation. For a set of random variables, the key difference is that Gram-Schmidt generates an uncorrelated basis that spans any *linear function* of the set of random variables. In contrast, conditional expectation (say $\mathbb{E}[X \mid Y, Z]$) gives the minimum variance estimate of X as an *arbitrary function* of Y and Z (e.g., a non-linear function). To highlight this, we note that Example 7.2 could be extended to include all integer powers of X. In contrast, Example 7.3 discusses an orthonormal basis not for Y, Z but for the indicator random variables of all outcomes for Y, Z.

8 Conclusion

We have reviewed vector space axioms, norms, inner products, the Cauchy–Schwarz inequality, and projections in $\mathbb{R}^n$. Then we discussed how these concepts connect to inner product spaces of random variables. Using this framework, one can unify projection and conditional expectation in probability spaces. This provides a solid foundation for further studies in estimation, signal processing, and machine learning.

Exercises

(1) Weighted inner product: projection and Gram–Schmidt in $\mathbb{C}^3$. Consider the weighted inner product on $\mathbb{C}^3$ defined by $\langle x, y\rangle_W = x^* W y$ with $W = \operatorname{diag}(2, 1, 3)$. Let $a = (1, i, 0)^\top$, $b = (1, 1, 1)^\top$, and $v = (2, 1, 0)^\top$.

   (a) Compute the orthogonal projection of v onto $\operatorname{span}\{a\}$ with respect to $\langle \cdot, \cdot \rangle_W$.

   (b) Apply Gram–Schmidt (with respect to $\langle \cdot, \cdot \rangle_W$) to the ordered set (a, b) to produce an orthonormal set (w_1, w_2).

   Solution: (a) With the linear in the first argument convention, the projection coefficient is $c = \frac{\langle v, a\rangle_W}{\langle a, a\rangle_W}$ and $P(v) = c\, a$. Compute $\langle a, a\rangle_W = a^* W a = (1, -i, 0) \cdot (2, i, 0) = 2 + 1 = 3$ and $\langle v, a\rangle_W = v^* W a = (2, 1, 0) \cdot (2, i, 0) = 4 + i$. Hence $P(v) = \frac{4+i}{3}\, a$.

   (b) Set $w_1 = a/\|a\|_W$ with $\|a\|_W = \sqrt{3}$, so $w_1 = a/\sqrt{3}$. Compute $\langle b, a\rangle_W = b^* W a = (1, 1, 1) \cdot (2, i, 0) = 2 + i$. Then $u_2 = b - \langle b, w_1\rangle_W w_1 = b - \frac{2+i}{\sqrt{3}} \cdot \frac{a}{\sqrt{3}} = b - \frac{2+i}{3}\, a = \big(\frac{1-i}{3},\ \frac{4-2i}{3},\ 1\big)^\top$. Its squared norm is $\|u_2\|_W^2 = u_2^* W u_2 = \frac{4}{9} + \frac{20}{9} + 3 = \frac{17}{3}$. Thus $\|u_2\|_W = \sqrt{17/3}$ and $w_2 = u_2/\|u_2\|_W$. The orthonormal set is $w_1 = \frac{1}{\sqrt{3}}(1, i, 0)^\top$ and $w_2 = \sqrt{\frac{3}{17}}\big(\frac{1-i}{3}, \frac{4-2i}{3}, 1\big)^\top$.

(2) Correlation bound and equality condition. Let X, Y be real random variables with finite second moments and $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$. Define the correlation $\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$. Show that $|\rho| \le 1$ and determine when equality holds.

Solution: In the inner-product space with $\langle U, V\rangle = \mathbb{E}[UV]$, apply Cauchy–Schwarz to the centered variables: $|\mathrm{Cov}(X,Y)| = |\langle X - \mathbb{E}[X], Y - \mathbb{E}[Y]\rangle| \le \|X - \mathbb{E}[X]\|\,\|Y - \mathbb{E}[Y]\| = \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$. Dividing yields $|\rho| \le 1$. Equality in Cauchy–Schwarz holds iff $Y - \mathbb{E}[Y] = c\,(X - \mathbb{E}[X])$ almost surely for some real c, i.e., iff Y is an affine function of X a.s.

(3) Two ways to project in $\mathbb{R}^3$. Let $a_1 = (1,2,2)^\top$, $a_2 = (0,1,1)^\top$, $W = \mathrm{span}\{a_1, a_2\}$, and $v = (1,0,2)^\top$.

(a) Use Gram–Schmidt to find an orthonormal basis of W and then compute $P_W(v)$.

(b) Use the normal equations with $A = [\,a_1\ a_2\,]$ to compute $P_W(v) = A\hat{x}$, where $\hat{x}$ solves $A^\top A\hat{x} = A^\top v$. Verify both answers agree and find the residual.

Solution: (a) $\|a_1\| = 3$, so $w_1 = a_1/3$. Then $\langle a_2, w_1\rangle = \frac{4}{3}$ and $u_2 = a_2 - \frac{4}{3}w_1 = a_2 - \frac{4}{9}a_1 = (-\frac{4}{9}, \frac{1}{9}, \frac{1}{9})^\top$. Its norm is $\|u_2\| = \sqrt{2}/3$, hence $w_2 = u_2/\|u_2\| = \frac{1}{3\sqrt{2}}(-4,1,1)^\top$. Now $\langle v, w_1\rangle = \frac{5}{3}$ and $\langle v, w_2\rangle = -\frac{2}{3\sqrt{2}}$, so $P_W(v) = \langle v, w_1\rangle w_1 + \langle v, w_2\rangle w_2 = \frac{5}{9}a_1 - \frac{1}{9}(-4,1,1)^\top = (1,1,1)^\top$.

(b) $A^\top A = \begin{bmatrix} 9 & 4 \\ 4 & 2 \end{bmatrix}$ and $A^\top v = (5,2)^\top$. Solving gives $\hat{x} = (1,-1)^\top$. Then $P_W(v) = A\hat{x} = a_1 - a_2 = (1,1,1)^\top$, matching (a). Residual $r = v - P_W(v) = (0,-1,1)^\top$ satisfies $A^\top r = 0$.

(4) Least-squares fit of a line (normal equations). Let $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \end{bmatrix}$. Find $\hat{x} = (c, m)$ that minimizes $\|Ax - b\|_2$ and compute $P_{\mathrm{range}(A)}(b) = A\hat{x}$. Verify orthogonality of the residual to the columns of A.

Solution: $A^\top A = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}$ and $A^\top b = (9, 18)^\top$. Solve $\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}\begin{bmatrix} c \\ m \end{bmatrix} = \begin{bmatrix} 9 \\ 18 \end{bmatrix}$: from $4c + 6m = 9$, $6c + 14m = 18$ we obtain $m = 0.9$, $c = 0.9$. Thus $\hat{x} = (0.9, 0.9)$ and $A\hat{x} = (0.9, 1.8, 2.7, 3.6)^\top$. Residual $r = b - A\hat{x} = (0.1, 0.2, -0.7, 0.4)^\top$ satisfies $A^\top r = (\sum r_i, \sum i\, r_i) = (0,0)$, hence $r \perp \mathrm{range}(A)$.

(5) Gramian, PSD, and linear dependence. Let $w_1 = (1,1,0)^\top$, $w_2 = (1,-1,0)^\top$, $w_3 = (2,0,0)^\top$ in $\mathbb{R}^3$. Form the Gramian $G = [g_{ij}]$ with $g_{ij} = \langle w_j, w_i\rangle$ under the standard inner product. Show $G \succeq 0$ and that G is not positive definite. Identify a nonzero v with $v^\top G v = 0$.

Solution: Compute inner products: $\langle w_1, w_1\rangle = 2$, $\langle w_2, w_2\rangle = 2$, $\langle w_3, w_3\rangle = 4$, $\langle w_2, w_1\rangle = 0$, $\langle w_3, w_1\rangle = 2$, $\langle w_3, w_2\rangle = 2$. Thus $G = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & 2 \\ 2 & 2 & 4 \end{bmatrix}$. For any $v = (v_1, v_2, v_3)^\top$, $v^\top G v = \left\| \sum_{i=1}^3 v_i w_i \right\|^2 \ge 0$, so $G \succeq 0$. Since $w_3 = w_1 + w_2$, choose $v = (1,1,-1)^\top$ to get $\sum v_i w_i = 0$, hence $v^\top G v = 0$ and G is not positive definite.

(6) Best affine estimator of a random variable. Let (X_1, X_2, Y) be real random variables with $\mu = \mathbb{E}[(X_1, X_2)^\top] = (1,-1)^\top$, $\Sigma = \mathrm{Cov}((X_1, X_2)^\top) = \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}$, $p = \mathrm{Cov}((X_1, X_2)^\top, Y) = (2,-1)^\top$,

and $\mathrm{Var}(Y) = 5$. Find the coefficients (c, w) of the best affine estimator $\widehat{Y} = c + w^\top (X_1, X_2)^\top$ and the minimum MSE.

Solution: Normal equations give $\Sigma w = p$ and $c = \mathbb{E}[Y] - w^\top \mu$ with $\mathbb{E}[Y] = 0$ (by assumption). Compute $\Sigma^{-1} = \frac{1}{15}\begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix}$, hence $w = \Sigma^{-1} p = \frac{1}{15}(9, -6)^\top = (0.6, -0.4)^\top$. Then $c = -w^\top \mu = -(0.6 \cdot 1 + (-0.4) \cdot (-1)) = -1$. Thus $\widehat{Y} = -1 + 0.6X_1 - 0.4X_2$. The minimum MSE equals $\mathrm{Var}(Y) - p^\top \Sigma^{-1} p = 5 - [2, -1] \cdot (0.6, -0.4)^\top = 5 - 1.6 = 3.4$.

(7) Law of total variance (numerical verification). Let $Y \in \{0, 1\}$ with $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = \frac{1}{2}$. Let $X \mid Y = 0$ take values 0 and 2 with equal probabilities, and $X \mid Y = 1 \equiv 1$ almost surely. Compute $\mathrm{Var}(X)$, $\mathbb{E}[\mathrm{Var}(X \mid Y)]$, and $\mathrm{Var}(\mathbb{E}[X \mid Y])$ and verify the law of total variance.

Solution: $X \mid Y = 0$: $\mathbb{E}[X \mid Y = 0] = 1$, $\mathrm{Var}(X \mid Y = 0) = 1$. $X \mid Y = 1$: $\mathbb{E}[X \mid Y = 1] = 1$, $\mathrm{Var}(X \mid Y = 1) = 0$. Hence $\mathbb{E}[\mathrm{Var}(X \mid Y)] = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2}$ and $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = 1$, so $\mathrm{Var}(\mathbb{E}[X \mid Y]) = 0$. Therefore $\mathrm{Var}(X) = \frac{1}{2} + 0 = \frac{1}{2}$, which also follows directly from the unconditional pmf.

(8) Gaussian MMSE for an additive noise channel. Let $X \sim \mathcal{N}(0, \sigma_X^2)$, $N \sim \mathcal{N}(0, \sigma_N^2)$ independent, and observe $Y = X + N$. Find the MMSE estimator $g(Y) = \mathbb{E}[X \mid Y]$ and the MMSE value.

Solution: (X, Y) is jointly Gaussian with $\mathrm{Cov}(X, Y) = \sigma_X^2$ and $\mathrm{Var}(Y) = \sigma_X^2 + \sigma_N^2$. For jointly Gaussian variables, $\mathbb{E}[X \mid Y]$ is linear: $g(Y) = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} Y = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2} Y$. The MMSE equals $\mathbb{E}[(X - g(Y))^2] = \sigma_X^2 - \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(Y)} = \sigma_X^2 - \frac{\sigma_X^4}{\sigma_X^2 + \sigma_N^2} = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}$.

(9) Conditional expectation and MMSE for a discrete pair. Let $Y \in \{0, 1\}$ with $\mathbb{P}(Y = 0) = 0.6$, $\mathbb{P}(Y = 1) = 0.4$. Conditional on Y, $X \mid Y = 0$ equals 0 with prob. 0.75 and 2 with prob. 0.25; $X \mid Y = 1$ equals 0 with prob. 0.25 and 2 with prob. 0.75. Compute $\mathbb{E}[X \mid Y]$, $\mathrm{Var}(X \mid Y)$, $\mathrm{mmse}(X \mid Y)$, and verify the law of total variance.

Solution: $\mathbb{E}[X \mid Y = 0] = 0 \cdot 0.75 + 2 \cdot 0.25 = 0.5$, $\mathbb{E}[X^2 \mid Y = 0] = 4 \cdot 0.25 = 1$, so $\mathrm{Var}(X \mid Y = 0) = 1 - 0.25 = 0.75$. $\mathbb{E}[X \mid Y = 1] = 2 \cdot 0.75 = 1.5$, $\mathbb{E}[X^2 \mid Y = 1] = 4 \cdot 0.75 = 3$, so $\mathrm{Var}(X \mid Y = 1) = 3 - 2.25 = 0.75$. Thus $\mathrm{mmse}(X \mid Y) = \mathbb{E}[\mathrm{Var}(X \mid Y)] = 0.6 \cdot 0.75 + 0.4 \cdot 0.75 = 0.75$. Furthermore, $\mathbb{E}[X] = 0.6 \cdot 0.5 + 0.4 \cdot 1.5 = 0.9$ and $\mathrm{Var}(\mathbb{E}[X \mid Y]) = \mathrm{Var}(\{0.5, 1.5\}) = 1.05 - 0.9^2 = 0.24$. Therefore, by the law of total variance, $\mathrm{Var}(X) = 0.75 + 0.24 = 0.99$, which agrees with the unconditional calculation $E[X^2] = 4 \cdot 0.45 = 1.8$, so $\mathrm{Var}(X) = 1.8 - 0.9^2 = 0.99$.

(10) Orthogonal complement and projection matrix in $\mathbb{R}^4$. Let $w_1 = (1, 1, 0, 0)^\top$, $w_2 = (0, 1, 1, 0)^\top$, $W = \mathrm{span}\{w_1, w_2\}$.

(a) Find a basis for $W^\perp$.

(b) Compute the orthogonal projector P_W via $P_W = A(A^\top A)^{-1} A^\top$, where $A = [\, w_1 \; w_2 \,]$.

(c) Compute $P_W(v)$ for $v = (1, 2, 3, 4)^\top$ and verify $v - P_W(v) \in W^\perp$.

Solution: (a) Solve $x \cdot w_1 = 0$ and $x \cdot w_2 = 0$: $x_1 + x_2 = 0$, $x_2 + x_3 = 0$. A basis is $\{(1, -1, 1, 0)^\top, (0, 0, 0, 1)^\top\}$.

(b) $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$, $A^\top A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, $(A^\top A)^{-1} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$. Hence $P_W = \frac{1}{3}\begin{bmatrix} 2 & 1 & -1 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$.

(c) $P_W(1,2,3,4)^\top = \frac{1}{3}(1,8,7,0)^\top$. Residual $r = (1,2,3,4)^\top - \frac{1}{3}(1,8,7,0)^\top = (\frac{2}{3}, -\frac{2}{3}, \frac{2}{3}, 4)^\top$ satisfies $r \cdot w_1 = 0$ and $r \cdot w_2 = 0$, so $r \in W^\perp$.

(11) Characterizing equality in Cauchy–Schwarz via projections. Let u, v be nonzero vectors in an inner-product space. Show that $|\langle u, v\rangle| = \|u\|\|v\|$ iff $P_{\mathrm{span}\{u\}}(v) = \alpha u$ for some α with $v - \alpha u = 0$ (i.e., v lies in the span of u). Interpret this in the space of random variables.

Solution: By the Projection Theorem, $v = P_{\mathrm{span}\{u\}}(v) + r$ with $r \perp u$. By the Pythagorean Theorem, $\|v\|^2 = \|P_{\mathrm{span}\{u\}}(v)\|^2 + \|r\|^2$. Also, $P_{\mathrm{span}\{u\}}(v) = \frac{\langle v,u\rangle}{\langle u,u\rangle} u$. Then $|\langle u, v\rangle| = \|u\|\|v\|$ iff $\|r\| = 0$ iff v is a scalar multiple of u. In the random-variable space with $\langle X, Y\rangle = \mathbb{E}[XY]$, equality holds iff $Y = cX$ almost surely for some c.

(12) Gram–Schmidt on a finite probability space. Let X take values in $\{-1, 0, 1\}$ with $\mathbb{P}(X = -1) = \mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - 2p$ for $p \in (0, \frac{1}{2}]$. Apply Gram–Schmidt in the inner-product space of real random variables (with $\langle U, V\rangle = \mathbb{E}[UV]$) to the list $(1, X, X^2)$ to produce an orthonormal set (Y_0, Y_1, Y_2).

Solution: First $Y_0 = 1$ since $\|1\|^2 = \mathbb{E}[1] = 1$. Center X: $\mathbb{E}[X] = (-1)p + 0\cdot(1-2p) + 1\cdot p = 0$, so X already has zero mean. Its norm is $\|X\|^2 = \mathbb{E}[X^2] = (-1)^2 p + 0 + 1^2 p = 2p$. Thus $Y_1 = X/\sqrt{2p}$.

Next orthogonalize X^2 against $\{1, Y_1\}$. Compute $\mathbb{E}[X^2] = 2p$ and $\mathbb{E}[X^3] = (-1)^3 p + 1^3 p = 0$, so $\langle X^2, Y_1\rangle = \mathbb{E}[X^2 \cdot X]/\sqrt{2p} = \mathbb{E}[X^3]/\sqrt{2p} = 0$. Hence $U_2 = X^2 - \langle X^2, 1\rangle \cdot 1 = X^2 - 2p$ and $\|U_2\|^2 = \mathbb{E}[(X^2 - 2p)^2] = \mathbb{E}[X^4] - 4p\,\mathbb{E}[X^2] + 4p^2 = (p + p) - 8p^2 + 4p^2 = 2p - 4p^2 = 2p(1 - 2p)$, using $X^4 = X^2$ on $\{-1, 0, 1\}$. Therefore $Y_2 = \dfrac{X^2 - 2p}{\sqrt{2p(1-2p)}}$. The set (Y_0, Y_1, Y_2) is orthonormal by construction.