Chapter 4

Representation and Approximation

4.1 Best Approximation

Suppose W is a subspace of a Banach space V. For any $\underline{v} \in V$, consider the problem of finding a vector $\underline{w} \in W$ such that $\|\underline{v} - \underline{w}\|$ is as small as possible.

Definition 4.1.1. *The vector $\underline{w} \in W$ is a **best approximation** of $\underline{v} \in V$ by vectors in W if*

$$\|\underline{v} - \underline{w}\| \leq \|\underline{v} - \underline{w}'\|$$

for all $\underline{w}' \in W$.

If W is spanned by the vectors $\underline{w}_1, \ldots, \underline{w}_n \in V$, then we can write

$$\begin{aligned} \underline{v} &= \underline{w} + \underline{e} \\ &= s_1\underline{w}_1 + \cdots + s_n\underline{w}_n + \underline{e}, \end{aligned}$$

where $\underline{e}$ is the approximation error.

This problem is, in general, rather difficult. However, if the norm $\|\cdot\|$ corresponds to the induced norm of an inner product, then one can use orthogonal projection and the problem is greatly simplified. This chapter focuses mainly on computing the best approximation of arbitrary vectors in a Hilbert space.

Theorem 4.1.2. *Suppose W is a subspace of a Hilbert space V and $\underline{v}$ is a vector in V. Then, we have the following:*

1. *The vector $\underline{w} \in W$ is the best approximation of $\underline{v} \in V$ by vectors in W if and only if $\underline{v} - \underline{w}$ is orthogonal to every vector in W.*

2. *If a best approximation of $\underline{v} \in V$ by vectors in W exists, it is unique.*

3. *If W has a countable orthogonal basis $\underline{w}_1, \underline{w}_2, \ldots$ and is closed, then*

$$\underline{w} = \sum_{i=1}^{\infty} \frac{\langle \underline{v} | \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i \tag{4.1}$$

is the best approximation of $\underline{v}$ by vectors in W.

Proof. Let $\underline{w} \in W$ and suppose $\underline{v} - \underline{w}$ is orthogonal to every vector in W. Let $\underline{w}' \in W$ such that $\underline{w}' \neq \underline{w}$. Then $\underline{v} - \underline{w}' = (\underline{v} - \underline{w}) + (\underline{w} - \underline{w}')$ and

$$\begin{aligned} \|\underline{v} - \underline{w}'\|^2 &= \|\underline{v} - \underline{w}\|^2 + 2\mathrm{Re}\,\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle + \|\underline{w} - \underline{w}'\|^2 \\ &= \|\underline{v} - \underline{w}\|^2 + \|\underline{w} - \underline{w}'\|^2 \\ &\geq \|\underline{v} - \underline{w}\|^2 . \end{aligned} \tag{4.2}$$

Conversely, suppose that $\|\underline{v} - \underline{w}'\| \geq \|\underline{v} - \underline{w}\|$ for every $\underline{w}' \in W$. From (4.2), we get

$$2\mathrm{Re}\,\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle + \|\underline{w} - \underline{w}'\|^2 \geq 0$$

for all $\underline{w}' \in W$. Note that every vector in W can be expressed as $\underline{w} - \underline{w}'$ where $\underline{w}' \in W$, it follows that

$$2\mathrm{Re}\,\langle \underline{v} - \underline{w} | \underline{w}'' \rangle + \|\underline{w}''\|^2 \geq 0 \tag{4.3}$$

for every $\underline{w}'' \in W$. If $\underline{w}'$ is in W and $\underline{w}' \neq \underline{w}$ then we may take

$$\underline{w}'' = -\frac{\langle \underline{w} - \underline{w}' | \underline{v} - \underline{w} \rangle}{\|\underline{w} - \underline{w}'\|^2} (\underline{w} - \underline{w}') .$$

Inequality (4.3) then reduces to the statement

$$-2\frac{|\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle|^2}{\|\underline{w} - \underline{w}'\|^2} + \frac{|\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle|^2}{\|\underline{w} - \underline{w}'\|^2} \geq 0.$$

This inequality holds if and only if $\langle \underline{v} - \underline{w} | \underline{w} - \underline{w}' \rangle = 0$. Therefore, $\underline{v} - \underline{w}$ is orthogonal to every vector in W. Hence the vector $\underline{w} \in W$ is a best approximation of $\underline{v} \in V$ by vectors in W if and only if $\underline{v} - \underline{w}$ is orthogonal to every vector in W.

Suppose $\underline{w}, \underline{w}' \in W$ are best approximations of $\underline{v}$ by vectors in W. Then $\|\underline{v} - \underline{w}\| = \|\underline{v} - \underline{w}'\|$ and (4.2) implies that $\|\underline{w} - \underline{w}'\| = 0$. That is, if a best approximation exists then it is unique.

Assume that W is closed and has a countable orthogonal basis $\underline{w}_1, \ldots, \underline{w}_n$ and let $\underline{w}$ be defined by (4.1). Then $\underline{v} - \underline{w}$ is orthogonal to $\underline{w}_j$ for $j \in \mathbb{N}$, i.e.,

$$\begin{aligned}\left\langle \underline{v} - \underline{w} | \underline{w}_j \right\rangle &= \left\langle \underline{v} | \underline{w}_j \right\rangle - \left\langle \sum_{i=1}^{n} \frac{\langle \underline{v} | \underline{w}_i \rangle}{\|\underline{w}_i\|^2} \underline{w}_i \middle| \underline{w}_j \right\rangle \\ &= \left\langle \underline{v} | \underline{w}_j \right\rangle - \frac{\langle \underline{v} | \underline{w}_j \rangle}{\|\underline{w}_i\|^2} \left\langle \underline{w}_j | \underline{w}_j \right\rangle = 0.\end{aligned}$$

That is, $\underline{v} - \underline{w}$ is orthogonal to every vector in W and therefore $\underline{w}$ is the best approximation of $\underline{v}$ by vectors in W. □

Definition 4.1.3. *Whenever the vector $\underline{w}$ in Theorem 4.1.2 exists, it is called the* ***orthogonal projection*** *of $\underline{v}$ onto W. If every vector in V has an orthogonal projection onto W, then the mapping $E : V \to W$, which assigns to each vector in V its orthogonal projection onto W, is called the orthogonal projection of V onto W.*

One can use Theorem 4.1.5 to verify that this is consistent with the concept of orthogonal projection from Definition 3.8.17.

P 4.1.4. *Let W be the subspace of $\mathbb{R}^2$ spanned by the vector $(1, 2)$. Using the standard inner product, let E be the orthogonal projection of $\mathbb{R}^2$ onto W. Find*

1. *a formula for $E(x_1, x_2)$*
2. *the matrix of E in the standard ordered basis, i.e., $E(x_1, x_2) = E\underline{x}$*
3. *$W^\perp$*
4. *an orthonormal basis in which E is represented by the matrix*

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Theorem 4.1.5. *Suppose W is a closed subspace of a separable Hilbert space V and let E denote the orthogonal projection of V on W. Then, E is an idempotent linear transformation of V onto W, $E\underline{w}' = \underline{0}$ iff $\underline{w}' \in W^\perp$, and*

$$V = W \oplus W^\perp.$$

Proof. Let $\underline{v}$ be any vector in V. Since $E\underline{v}$ is the best approximation of $\underline{v}$ by vectors in W, it follows that $\underline{v} \in W$ implies $E\underline{v} = \underline{v}$. Therefore, $E\left(E\underline{v}\right) = E\underline{v}$ for any $\underline{v} \in V$ since $E\underline{v} \in W$. That is, $E^2 = E$ and E is idempotent.

To show that E is a linear transformation, let $\underline{w}_1, \underline{w}_2, \ldots$ be a countable orthonormal basis for W (whose existence follows from Theorem 3.7.16). Using part 3 of Theorem 4.1.2, we find that

$$\begin{aligned} E\left(s_1\underline{v}_1 + \underline{v}_2\right) &= \sum_{i=1}^{\infty} \langle s_1\underline{v}_1 + \underline{v}_2 | \underline{w}_i \rangle \underline{w}_i \\ &= s_1 \sum_{i=1}^{\infty} \langle \underline{v}_1 | \underline{w}_i \rangle \underline{w}_i + \sum_{i=1}^{\infty} \langle \underline{v}_2 | \underline{w}_i \rangle \underline{w}_i \\ &= s_1 E\underline{v}_1 + E\underline{v}_2. \end{aligned}$$

Therefore, E is a linear transformation. It also follows that $E\underline{w}' = \underline{0}$ iff $\underline{w}' \in W^{\perp}$ because $W^{\perp}$ can be defined by the fact that $\langle \underline{w}' | \underline{w}_i \rangle = 0$ for $i \in \mathbb{N}$.

Again, let $\underline{v} \in V$ and recall that (by Theorem 4.1.2) $E\underline{v}$ is the unique vector in W such that $\underline{v} - E\underline{v}$ is in $W^{\perp}$. Therefore, the equation $\underline{v} = E\underline{v} + (\underline{v} - E\underline{v})$ gives a unique decomposition of $\underline{v}$ into $E\underline{v} \in W$ and $\underline{v} - E\underline{v} \in W^{\perp}$. This unique decomposition implies that V is the direct sum of W and $W^{\perp}$. Lastly, one finds from the definition of $W^{\perp}$ that

$$W \cap W^{\perp} = \{\underline{u} \in W | \langle \underline{u} | \underline{w} \rangle = 0 \; \forall \; \underline{w} \in W\} \subseteq \{\underline{u} \in W | \langle \underline{u} | \underline{u} \rangle = 0\} = \{\underline{0}\}.$$

□

Corollary 4.1.6. *Let W be a closed subspace of a separable Hilbert space V and E be the orthogonal projection of V on W. Then $I - E$ is the orthogonal projection of V on $W^{\perp}$.*

Proof. This follows directly from the orthogonal decomposition in Theorem 4.1.5. One can also verify that $I - E$ is an idempotent linear transformation of V with range $W^{\perp}$ and nullspace W. From Definition 3.8.17, we see that $I - E$ is an orthogonal projection. □

Theorem 4.1.5 also implies the following result, known as Bessel's inequality.

Corollary 4.1.7. *Let $\underline{v}_1, \underline{v}_2, \ldots$ be a countable orthogonal set of distinct non-zero vectors in an inner-product space V. If $\underline{v} \in V$ then*

$$\sum_{i=1}^{\infty} \frac{|\langle \underline{v}|\underline{v}_i\rangle|^2}{\|\underline{v}_i\|^2} \leq \|\underline{v}\|^2 .$$

Moreover, equality holds if and only if

$$\underline{v} = \sum_{i=1}^{\infty} \frac{\langle \underline{v}|\underline{v}_i\rangle}{\|\underline{v}_i\|^2} \underline{v}_i.$$

Proof. Define

$$\underline{w} = \sum_{i=1}^{\infty} \frac{\langle \underline{v}|\underline{v}_i\rangle}{\|\underline{v}_i\|^2} \underline{v}_i.$$

Then $\underline{v} = \underline{w} + \underline{u}$, where $\langle \underline{w}|\underline{u}\rangle = 0$ and $\|\underline{v}\|^2 = \|\underline{w}\|^2 + \|\underline{u}\|^2$. Noting that

$$\|\underline{w}\|^2 = \sum_{i=1}^{\infty} \frac{|\langle \underline{v}|\underline{v}_i\rangle|^2}{\|\underline{v}_i\|^2},$$

we obtain the desired result, with equality iff $\underline{u} = \underline{0}$. □

If $\underline{v}_1, \underline{v}_2, \ldots$ is an orthonormal set, Bessel's inequality states that

$$\sum_{i=1}^{\infty} |\langle \underline{v}|\underline{v}_i\rangle|^2 \leq \|\underline{v}\|^2 .$$

It follows that the vector $\underline{v}$ is in the closure of the subspace spanned by $\underline{v}_1, \underline{v}_2, \ldots$ if an only if

$$\underline{v} = \sum_{i=1}^{\infty} \langle \underline{v}|\underline{v}_i\rangle \, \underline{v}_i.$$

4.2 Computing Approximations in Hilbert Spaces

4.2.1 Normal Equations

Suppose V is a Hilbert space the subspace W is spanned by $\underline{w}_1, \ldots, \underline{w}_n \in V$. Consider the situation where the sequence $\underline{w}_1, \ldots, \underline{w}_n$ is linearly independent, but not orthogonal. In this case, it is not possible to apply (4.1) directly. It is nevertheless possible to obtain a similar expression for the best approximation of $\underline{v}$ by vectors in

W. Theorem 4.1.2 asserts that $\hat{\underline{v}} \in W$ is a best approximation of $\underline{v} \in V$ by vectors in W if and only if $\underline{v} - \hat{\underline{v}}$ is orthogonal to every vector in W. This implies that

$$\left\langle \underline{v} - \hat{\underline{v}} | \underline{w}_j \right\rangle = \left\langle \underline{v} - \sum_{i=1}^{n} s_i \underline{w}_i \middle| \underline{w}_j \right\rangle = 0$$

or, equivalently,

$$\sum_{i=1}^{n} s_i \left\langle \underline{w}_i | \underline{w}_j \right\rangle = \left\langle \underline{v} | \underline{w}_j \right\rangle$$

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

$$\begin{bmatrix} \langle \underline{w}_1 | \underline{w}_1 \rangle & \langle \underline{w}_2 | \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_1 \rangle \\ \langle \underline{w}_1 | \underline{w}_2 \rangle & \langle \underline{w}_2 | \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1 | \underline{w}_n \rangle & \langle \underline{w}_2 | \underline{w}_n \rangle & \cdots & \langle \underline{w}_n | \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} \langle \underline{v} | \underline{w}_1 \rangle \\ \langle \underline{v} | \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v} | \underline{w}_n \rangle \end{bmatrix}.$$

We can rewrite this matrix equation as

$$G\underline{s} = \underline{t}$$

where

$$\underline{t}^T = \left(\langle \underline{v} | \underline{w}_1 \rangle, \langle \underline{v} | \underline{w}_2 \rangle, \ldots, \langle \underline{v} | \underline{w}_n \rangle \right)$$

is the **cross-correlation vector**, and

$$\underline{s}^T = (s_1, s_2, \ldots, s_n)$$

is the vector of coefficients. Equations of this form are collectively known as the **normal equations**.

Definition 4.2.1. *The $n \times n$ matrix*

$$G = \begin{bmatrix} \langle \underline{w}_1 | \underline{w}_1 \rangle & \langle \underline{w}_2 | \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_1 \rangle \\ \langle \underline{w}_1 | \underline{w}_2 \rangle & \langle \underline{w}_2 | \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1 | \underline{w}_n \rangle & \langle \underline{w}_2 | \underline{w}_n \rangle & \cdots & \langle \underline{w}_n | \underline{w}_n \rangle \end{bmatrix} \tag{4.4}$$

*is called the **Grammian** matrix. Since $G_{ji} = \left\langle \underline{w}_i | \underline{w}_j \right\rangle$, it follows that the Grammian is a Hermitian symmetric matrix, i.e., $G^H = G$.*

Definition 4.2.2. *A matrix $M \in F^{n\times n}$ is **positive-semidefinite** if $\underline{v}^H M\underline{v} \geq 0$ for all $\underline{v} \in F^n$. A matrix $M \in F^{n\times n}$ is **positive-definite** if $\underline{v}^H M\underline{v} > 0$ for all $\underline{v} \in F^n - \{\underline{0}\}$.*

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

Theorem 4.2.3. *A Grammian matrix G is always positive-semidefinite. Furthermore, it is positive-definite if and only if the sequence of vectors $\underline{w}_1, \ldots, \underline{w}_n$ is linearly independent.*

Proof. Let $\underline{v} = (v_1, \ldots, v_n)^T \in F^n$. Then,

$$\begin{aligned}
\underline{v}^H G\underline{v} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{v}_j G_{ji} v_i = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{v}_j \left\langle \underline{w}_i | \underline{w}_j \right\rangle v_i \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \left\langle v_i\underline{w}_i | v_j\underline{w}_j \right\rangle = \left\langle \sum_{i=1}^{n} v_i\underline{w}_i \middle| \sum_{j=1}^{n} v_j\underline{w}_j \right\rangle \\
&= \left\| \sum_{i=1}^{n} v_i\underline{w}_i \right\|^2 \geq 0.
\end{aligned} \tag{4.5}$$

That is, $\underline{v}^H G\underline{v} \geq 0$ for all $\underline{v} \in F^n$.

Suppose that G is not positive-definite. Then, there exists $\underline{v} \in F^n - \{\underline{0}\}$ such that $\underline{v}^H G\underline{v} = 0$. By (4.5), this implies that

$$\sum_{i=1}^{n} v_i\underline{w}_i = 0$$

and hence the sequence of vectors $\underline{w}_1, \ldots, \underline{w}_n$ is not linearly independent.

Conversely, if G is positive-definite then $\underline{v}^H G\underline{v} > 0$ and

$$\left\| \sum_{i=1}^{n} v_i\underline{w}_i \right\| > 0$$

for all $\underline{v} \in F^n - \{\underline{0}\}$. Therefore, the sequence of vectors $\underline{w}_1, \ldots, \underline{w}_n$ is linearly independent. □

4.2.2 Orthogonality Principle

Theorem 4.2.4. *Let $\underline{w}_1, \ldots, \underline{w}_n$ be vectors in an inner-product space V and denote the span of $\underline{w}_1, \ldots, \underline{w}_n$ by W. For any vector $\underline{v} \in V$, the norm of the error vector*

$$\underline{e} = \underline{v} - \sum_{i=1}^{n} s_i \underline{w}_i \tag{4.6}$$

*is minimized when the error vector $\underline{e}$ is orthogonal to every vector in W. If $\hat{\underline{v}}$ denotes the **least-squares** approximation of $\underline{v}$ then*

$$\left\langle \underline{v} - \hat{\underline{v}} | \underline{w}_j \right\rangle = 0$$

for $j = 1, \ldots, n$.

Proof. Minimizing $\|\underline{e}\|^2$, where $\underline{e}$ is given by (4.6) requires minimizing

$$\begin{aligned} J(\underline{s}) &= \left\langle \underline{v} - \sum_{i=1}^{n} s_i \underline{w}_i \middle| \underline{v} - \sum_{j=1}^{n} s_j \underline{w}_j \right\rangle \\ &= \langle \underline{v} | \underline{v} \rangle - \sum_{i=1}^{n} \langle s_i \underline{w}_i | \underline{v} \rangle - \sum_{j=1}^{n} \left\langle \underline{v} | s_j \underline{w}_j \right\rangle + \sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle s_i \underline{w}_i | s_j \underline{w}_j \right\rangle \\ &= \langle \underline{v} | \underline{v} \rangle - \sum_{i=1}^{n} s_i \langle \underline{w}_i | \underline{v} \rangle - \sum_{j=1}^{n} \bar{s}_j \left\langle \underline{v} | \underline{w}_j \right\rangle + \sum_{i=1}^{n} \sum_{j=1}^{n} s_i \bar{s}_j \left\langle \underline{w}_i | \underline{w}_j \right\rangle . \end{aligned}$$

Taking the gradient of $J(\underline{s})$, we get

$$\begin{aligned} \nabla J(\underline{s}) &= - \begin{bmatrix} \langle \underline{v} | \underline{w}_1 \rangle \\ \langle \underline{v} | \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v} | \underline{w}_n \rangle \end{bmatrix} + \begin{bmatrix} \langle \underline{w}_1 | \underline{w}_1 \rangle & \langle \underline{w}_2 | \underline{w}_1 \rangle & \ldots & \langle \underline{w}_n | \underline{w}_1 \rangle \\ \langle \underline{w}_1 | \underline{w}_2 \rangle & \langle \underline{w}_2 | \underline{w}_2 \rangle & \ldots & \langle \underline{w}_n | \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1 | \underline{w}_n \rangle & \langle \underline{w}_2 | \underline{w}_n \rangle & \ldots & \langle \underline{w}_n | \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} \\ &= \underline{0}. \end{aligned}$$

In matrix form, this yields the familiar equation

$$G\underline{s} = \underline{t}.$$

To ensure that this extremum is in fact a minimum, we compute the Hessian matrix

$$\nabla^2 J(\underline{s}) = G.$$

Since G is a positive-semidefinite matrix, the extremum is indeed a minimum.

This implies that $\|\underline{e}\|^2$ is minimized if and only if $G\underline{s} = \underline{t}$. That is, $\|\underline{e}\|^2$ is minimized if and only if $\underline{v} - \hat{\underline{v}}$ is orthogonal to every vector in W. □

Note that it is also possible to prove this theorem using the Cauchy-Schwarz inequality or the projection theorem.

4.3 Approximation for Systems of Linear Equations

4.3.1 Matrix Representation

For finite-dimensional vector spaces, least-squares (i.e., best approximation) problems have natural matrix representations. Suppose $V = F^m$ and $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n \in V$ are column vectors. Then, the approximation vector is given by

$$\underline{\hat{v}} = \sum_{i=1}^{n} s_i \underline{w}_i$$

In matrix form, we have

$$\underline{\hat{v}} = A\underline{s},$$

where $A = [\underline{w}_1 \cdots \underline{w}_n]$. The optimization problem can then be reformulated as follows. Determine $\underline{s} \in F^n$ such that

$$\|\underline{e}\|^2 = \|\underline{v} - \underline{\hat{v}}\|^2 = \|\underline{v} - A\underline{s}\|^2$$

is minimized. Note that this occurs when the error vector is orthogonal to every vector in W, i.e.,

$$\langle \underline{e} | \underline{w}_j \rangle = \langle \underline{v} - \underline{\hat{v}} | \underline{w}_j \rangle = \langle \underline{v} - A\underline{s} | \underline{w}_j \rangle = 0$$

for $j = 1, \ldots, n$.

4.3.2 Standard Inner Products

When $\|\cdot\|$ is the norm induced by the standard inner product, these conditions can be expressed as

$$\begin{bmatrix} \underline{w}_1^H \\ \vdots \\ \underline{w}_n^H \end{bmatrix} (\underline{v} - A\underline{s}) = \underline{0}.$$

Using the definition of A, we obtain

$$A^H A\underline{s} = A^H \underline{v}.$$

The matrix $A^H A$ is the Grammian G defined in (4.4). The vector $A^H \underline{v}$ is the cross correlation vector $\underline{t}$.

When the vectors $\underline{w}_1, \ldots, \underline{w}_n$ are linearly independent, the Grammian matrix is positive definite and hence invertible. The optimal solution for the least-squares problem is therefore given by

$$\underline{s} = \left(A^H A\right)^{-1} A^H \underline{v} = G^{-1} \underline{t}.$$

The matrix $\left(A^H A\right)^{-1} A^H$ is often called the **pseudoinverse**.

The best approximation of $\underline{v} \in V$ by vectors in W is equal to

$$\hat{\underline{v}} = A\underline{s} = A\left(A^H A\right)^{-1} A^H \underline{v}.$$

The matrix $P = A\left(A^H A\right)^{-1} A^H$ is called the **projection matrix** for the range of A. It defines an orthogonal projection onto the range of A (i.e., the subspace spanned by the columns of A).

4.3.3 Generalized Inner Products

We can also consider the case of a general inner product. Recall that an inner product on V is completely determined by the values

$$h_{ji} = \left\langle \underline{e}_i | \underline{e}_j \right\rangle,$$

and that this inner product can be expressed as

$$\langle \underline{v} | \underline{w} \rangle = \underline{w}^H H \underline{v}.$$

Minimizing $\|\underline{e}\|^2 = \|\underline{v} - A\underline{s}\|^2$ and using the orthogonality principle lead to the matrix equation

$$A^H H A \underline{s} = A^H H \underline{v}.$$

When the vectors $\underline{w}_1, \ldots, \underline{w}_n$ are linearly independent, the optimal solution is given by

$$\underline{s} = \left(A^H H A\right)^{-1} A^H H \underline{v}.$$

4.3.4 Minimum Error

Let $\underline{\hat{v}} \in W$ be the best approximation of $\underline{v}$ by vectors in W. Again, we can write

$$\underline{v} = \underline{\hat{v}} + \underline{e}$$

where $\underline{e} \in W^{\perp}$ is the minimum achievable error. The squared norm of the minimum error is given implicitly by

$$\|\underline{v}\|^2 = \|\underline{\hat{v}} + \underline{e}\|^2 = \langle \underline{\hat{v}} + \underline{e} | \underline{\hat{v}} + \underline{e} \rangle = \langle \underline{\hat{v}} | \underline{\hat{v}} \rangle + \langle \underline{e} | \underline{e} \rangle = \|\underline{\hat{v}}\|^2 + \|\underline{e}\|^2 .$$

We can then find an explicit expression for the approximation error,

$$\begin{aligned}
\|\underline{e}\|^2 = \|\underline{v}\|^2 - \|\underline{\hat{v}}\|^2 &= \underline{v}^H H \underline{v} - \underline{\hat{v}}^H H \underline{\hat{v}} \\
&= \underline{v}^H H \underline{v} - \underline{s}^H A^H H A \underline{s} \\
&= \underline{v}^H H \underline{v} - \underline{v}^H H A \left(A^H H A\right)^{-1} A^H H \underline{v} \\
&= \underline{v}^H \left(H - H A \left(A^H H A\right)^{-1} A^H H\right) \underline{v}.
\end{aligned}$$

4.4 Applications and Examples in Signal Processing

4.4.1 Linear Regression

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a collection of points in $\mathbb{R}^2$. A **linear regression** problem consists in finding scalars a and b such that

$$y_i \approx a x_i + b$$

for $i = 1, \ldots, n$. Definite the error component e_i by $e_i = y_i - a x_i - b$, then

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = a \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + b \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}.$$

In vector form, we can rewrite this equation as

$$\underline{y} = A \underline{s} + \underline{e},$$

where $\underline{y} = (y_1, \ldots, y_n)^T$, $\underline{s} = (a, b)^T$, $\underline{e} = (e_1, \ldots, e_n)^T$, and

$$A = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix}.$$

This equation has a form analog to the matrix representation of a least-squares problems. Consider the goal of minimizing $\|\underline{e}\|^2$. The line that minimizes the sums of the squares of the *vertical* distances between the data abscissas and the line is then given by

$$\underline{s} = \left(A^H A\right)^{-1} A^H \underline{y}.$$

4.4.2 Linear Minimum Mean-Squared Estimation

Let $Y, X_1, \ldots, X_n$ be a set of zero-mean random variables. The goal of the linear minimum mean-squared estimation (LMMSE) problem is to find coefficients $s_1, \ldots, s_n$ such that

$$\hat{Y} = s_1 X_1 + \cdots + s_n X_n.$$

minimizes the MSE $\mathrm{E}[|Y - \hat{Y}|^2]$. Using the inner product defined by

$$\langle X|Y\rangle = \mathrm{E}\left[X\overline{Y}\right], \tag{4.7}$$

we can compute the linear minimum mean-squared estimate $\hat{Y}$ using

$$G\underline{s} = \underline{t},$$

where

$$G = \begin{bmatrix} \mathrm{E}\left[X_1\overline{X}_1\right] & \mathrm{E}\left[X_2\overline{X}_1\right] & \cdots & \mathrm{E}\left[X_n\overline{X}_1\right] \\ \mathrm{E}\left[X_1\overline{X}_2\right] & \mathrm{E}\left[X_2\overline{X}_2\right] & \cdots & \mathrm{E}\left[X_n\overline{X}_2\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}\left[X_1\overline{X}_n\right] & \mathrm{E}\left[X_2\overline{X}_n\right] & \cdots & \mathrm{E}\left[X_n\overline{X}_n\right] \end{bmatrix}$$

and

$$\underline{t} = \begin{bmatrix} \mathrm{E}\left[Y\overline{X}_1\right] \\ \mathrm{E}\left[Y\overline{X}_2\right] \\ \vdots \\ \mathrm{E}\left[Y\overline{X}_n\right] \end{bmatrix}.$$

If the matrix G is invertible, the minimum mean-squared error is given by

$$\left\|Y - \hat{Y}\right\|^2 = \mathrm{E}\left[Y\overline{Y}\right] - \underline{t}^H G^{-1} \underline{t}.$$

4.4.3 The Wiener Filter

Suppose that the sequence of zero-mean random variables $\{X[t]\}$ is wide-sense stationary, and consider the FIR filter

$$\begin{aligned} Y[t] &= \sum_{k=0}^{K-1} h[k]X[t-k] \\ &= \begin{bmatrix} X[t] & \dots & X[t-K+1] \end{bmatrix} \begin{bmatrix} h[0] \\ \vdots \\ h[K-1] \end{bmatrix} = (\underline{X}[t])^T \underline{h}. \end{aligned}$$

The goal is to design this filter in such a way that its output is as close as possible to a desired sequence $\{Z[t]\}$. In particular, we want to minimize the mean-squared error

$$\|Z[t] - Y[t]\|^2 = \mathrm{E}\left[|Z[t] - Y[t]|^2\right].$$

By the orthogonality principle, the mean-squared error is minimized when the error is orthogonal to the data; that is, for $j = 0, 1, \dots, K-1$, we have

$$\left\langle Z[t] - \sum_{k=0}^{K-1} h[k]X[t-k] \middle| X[t-j] \right\rangle = 0,$$

or, equivalently, we can write

$$\langle Z[t]|X[t-j]\rangle = \sum_{k=0}^{K-1} h[k] \langle X[t-k]|X[t-j]\rangle.$$

Using (4.7), we obtain

$$\mathrm{E}\left[Z[t]\overline{X}[t-j]\right] = \sum_{k=0}^{K-1} h[k]\mathrm{E}\left[X[t-k]\overline{X}[t-j]\right]. \tag{4.8}$$

where $j = 1, \dots, K-1$.

For this specific case where the normal equations are defined in terms of the expectation operator, these equations are called the **Wiener-Hopf** equations. The Grammian of the Wiener-Hopf equations can be expressed in a more familiar form using the autocorrelation matrix. Recall that $\{X[t]\}$ is a wide-sense stationary process. As such, we have

$$R_{xx}(j-k) = R_{xx}(j,k) = \mathrm{E}\left[X[t-k]\overline{X}[t-j]\right] = \langle X[t-k]|X[t-j]\rangle.$$

Also define

$$R_{zx}(j) = \mathrm{E}\left[Z[t]\overline{X}[t-j]\right] = \langle Z[t]|X[t-j]\rangle\,.$$

Using this notation, we can rewrite (4.8) as

$$R_{zx} = \begin{bmatrix} R_{zx}(0) \\ R_{zx}(1) \\ \vdots \\ R_{zx}(K-1) \end{bmatrix} = R_{xx} \begin{bmatrix} h[0] \\ h[1] \\ \vdots \\ h[K-1] \end{bmatrix}$$

where the $K \times K$ autocorrelation matrix is given by

$$R_{xx} = \begin{bmatrix} R_{xx}[0] & \overline{R}_{xx}[1] & \cdots & \overline{R}_{xx}[K-1] \\ R_{xx}[1] & R_{xx}[0] & \cdots & \overline{R}_{xx}[K-2] \\ \vdots & \vdots & \ddots & \vdots \\ R_{xx}[K-1] & R_{xx}[K-2] & \cdots & R_{xx}[0] \end{bmatrix}.$$

Note that the matrix R_{xx} is Toeplitz, i.e., all the elements on a diagonal are equal. Assuming that R_{xx} is invertible, the optimal filter taps are then given by

$$\underline{h} = R_{xx}^{-1} R_{zx}.$$

The minimum mean-squared error is given by

$$\begin{aligned} \|Z-Y\|^2 &= \|Z\|^2 - \|Y\|^2 \\ &= \mathrm{E}[Z\overline{Z}] - \mathrm{E}\left[\underline{h}^H \underline{\overline{X}}\,\underline{X}^T \underline{h}\right] \\ &= \mathrm{E}[Z\overline{Z}] - \underline{h}^H R_{xx} \underline{h} \\ &= \mathrm{E}[Z\overline{Z}] - R_{zx}^H \underline{h}, \end{aligned}$$

where t can be ignored because the processes are WSS.

4.4.4 LMMSE Filtering in Practice

While theoretical treatments of optimal filtering often assume one has well-defined random variables with known statistics, this is rarely the case in practice. Yet, there is a very close connection between Wiener filtering and natural data driven approaches. Consider the problem from the previous section and let $x[1], x[2], \ldots_1, x[N]$ and $z[1], z[2], \ldots, z[N]$ be realizations of the random processes.

As an application, one can think of the $x[t]$ sequence as the received samples in a wireless communication system and the $z[t]$ sequence as a *pilot sequence* (i.e., known to both the transmitter and receiver). It is assumed the transmitted sequence has been convolved with an unknown LTI system. This type of degradation is known as intersymbol interference (ISI) and the goal is to find a linear filter $h[0], h[1], \ldots, h[K-1]$ that removes as much ISI as possible. A suitable cost function for this goal is

$$J(\underline{h}) = \sum_{t=K}^{N} \lambda^{N-t} \left| z[t] - \sum_{k=0}^{K-1} h[k]x[t-k] \right|^2,$$

where the exponential weighting factor λ emphasizes the most recently received symbols because, in reality, the channel conditions are changing with time.

Using the vector $\underline{z} = [\ z[K] \quad z[K+1] \quad \cdots \quad z[N]\ ]$ and the matrix

$$A = \begin{bmatrix} x[K] & x[K-1] & \cdots & x[1] \\ x[K+1] & x[K] & \cdots & x[2] \\ \vdots & \vdots & \ddots & \vdots \\ x[N] & x[N-1] & \cdots & x[N-K+1] \end{bmatrix},$$

we can rewrite this cost function as

$$J(\underline{h}) = (A\underline{h} - \underline{z})^H \Lambda (A\underline{h} - \underline{z}),$$

where Λ is a diagonal matrix whose diagonal contains $[\ \lambda^{N-K} \quad \lambda^{N-K+1} \quad \cdots \quad \lambda^1 \quad \lambda^0\ ]$. Using the orthogonality principle, one finds that the optimal solution is given by the normal equation

$$A^H \Lambda A \underline{h} = A^H \Lambda \underline{z}.$$

To see the connection with Wiener filtering, the key observation is that the matrix $A^H \Lambda A$ and the vector $A^H \Lambda \underline{z}$ are sample-average estimates of the correlation matrix and cross-correlation vector. This is because, for large N and λ close to 1, we have

$$\left[A^H \Lambda A\right]_{ij} = \sum_{t=K}^{N} \lambda^{N-t} x[t-j+1]\overline{x}[t-i+1] \approx \frac{R_{xx}(i-j)}{1-\lambda}$$

and

$$\left[A^H \Lambda \underline{z}\right]_i = \sum_{t=K}^{N} \lambda^{N-t} z[t]\overline{x}[t-i+1] \approx \frac{R_{zx}(i)}{1-\lambda}.$$

Another benefit of this approach is that, as each new sample arrives, the solution $\underline{h}$ can be updated with low complexity. Consider the matrix $G_N = A^H \Lambda A$ and vector $\underline{b}_N = A^H \Lambda \underline{z}$ as a function of N. Then, $G_{N+1} = \lambda G_N + \underline{u}^H \underline{u}$ and $\underline{t}_{N+1} = \lambda \underline{b}_N + z[N+1]\underline{u}^H$, where

$$\underline{u} = \begin{bmatrix} x[N+1] & x[N] & \cdots & x[N-K+2] \end{bmatrix}.$$

The updated solution vector $\underline{h}_{N+1} = G_{N+1}^{-1} \underline{b}_{N+1}$ can be computed efficiently using the Sherman-Morrison matrix inversion formula.

4.5 Dual Approximation

4.5.1 Minimum-Norm Solutions

In many cases, one is interested in finding the minimum-norm vector that satisfies some feasibility constraints. For example, an underdetermined system of linear equations has an infinite number of solutions. But, in practice, it often makes sense to prefer the minimum-norm solution over other solutions. Finding this solution is very similar to finding the best approximation.

Let V be a Hilbert space and $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$ be a set of linearly independent vectors in W. For any $\underline{v} \in V$, consider finding the scalars $s_1, s_2, \ldots, s_n$ that minimize

$$\left\| \underline{v} - \sum_{i=1}^{n} s_i \underline{w}_i \right\|.$$

The answer is clearly given by the best approximation of $\underline{v}$ by vectors in the span of $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$. The orthogonality principle tells us that $s_1, s_2, \ldots, s_n$ must satisfy

$$\begin{bmatrix} \langle \underline{w}_1 | \underline{w}_1 \rangle & \langle \underline{w}_2 | \underline{w}_1 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_1 \rangle \\ \langle \underline{w}_1 | \underline{w}_2 \rangle & \langle \underline{w}_2 | \underline{w}_2 \rangle & \cdots & \langle \underline{w}_n | \underline{w}_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \underline{w}_1 | \underline{w}_n \rangle & \langle \underline{w}_2 | \underline{w}_n \rangle & \cdots & \langle \underline{w}_n | \underline{w}_n \rangle \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{bmatrix} = \begin{bmatrix} \langle \underline{v} | \underline{w}_1 \rangle \\ \langle \underline{v} | \underline{w}_2 \rangle \\ \vdots \\ \langle \underline{v} | \underline{w}_n \rangle \end{bmatrix}. \tag{4.9}$$

The same problem can also be posed in a different manner.

Theorem 4.5.1. *Let V be a Hilbert space and $\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n$ be a set of linearly independent vectors in V. The* ***dual approximation*** *problem is to find the vector*

$\underline{w} \in V$ of minimum-norm that satisfes $\langle \underline{w} | \underline{w}_i \rangle = c_i$ for $i = 1, \ldots, n$. This vector is given by

$$\underline{w} = \sum_{i=1}^{n} s_i \underline{w}_i,$$

where the coefficients $s_1, s_2, \ldots, s_n$ can be found by solving (4.9) *with $\langle \underline{v} | \underline{w}_i \rangle = c_i$.*

Proof. Let $W = \operatorname{span}(\underline{w}_1, \underline{w}_2, \ldots, \underline{w}_n)$ and notice that the subset

$$A = \{\underline{u} \in V | \, \langle \underline{u} | \underline{w}_i \rangle = c_i \; \forall \; i = 1, \ldots, n\}$$

is simply the orthogonal complement $W^{\perp}$ translated by any vector $\underline{v} \in A$. Therefore, the vector achieving $\min_{\underline{u} \in A} \|\underline{u}\|$ is the error vector in the best approximation of some $\underline{v} \in A$ by vectors in $W^{\perp}$. Using the unique decomposition $\underline{v} = \hat{\underline{v}} + \underline{w}$ implied by the orthogonal decomposition $V = W^{\perp} \oplus W$, one finds that the error vector $\underline{w}$ must lie in W. Moreover, the normal equations, given by modifying (4.9), show that the error vector $\underline{w}$ is the unique vector in W that satisfies $\langle \underline{w} | \underline{w}_i \rangle = c_i$ for $i = 1, \ldots, n$. □

4.5.2 Underdetermined Linear Systems

Let $A \in \mathbb{C}^{m \times n}$ with $m < n$ be the matrix representation of an underdetermined system of linear equations and $\underline{v} \in \mathbb{C}^m$ be any column vector. Then, the dual approximation theorem can be applied to solve the problem

$$\min_{\underline{s}: A\underline{s} = \underline{v}} \|\underline{s}\|.$$

To see this as a dual approximation, we can rewrite the constraint as $(A^H)^H \underline{s} = \underline{v}$. Then, the theorem concludes that the minimum norm solution lies in $\mathcal{R}(A^H)$ (i.e., the column space of A^H). Using this, one can define $\hat{\underline{s}} = A^H \underline{t}$ and see that $A(A^H \underline{t}) = \underline{v}$. If the rows of A are linearly independent, then the columns of A^H are linearly independent and $(AA^H)^{-1}$ exists. In this case, the solution $\hat{\underline{s}}$ can be obtained in closed form and is given by

$$\hat{\underline{s}} = A^H \left(AA^H\right)^{-1} \underline{v}.$$