Problem Set 6 — *Due Friday, December 14, before class starts*

For the Exercise Sessions on Nov 30 and Dec 7

Last name	First name	SCIPER Nr	Points

Problem 1: Eckart–Young Theorem

In class, we proved the converse part of the Eckart–Young theorem for the spectral norm. Here, you do the same for the case of the Frobenius norm.

(a) For any matrix A of dimension $m \times n$ and an arbitrary orthonormal basis $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ of $\mathbb{C}^n$, prove that

$$\|A\|_F^2 \quad = \quad \sum_{k=1}^{n} \|A\mathbf{x}_k\|^2. \tag{1}$$

(b) Consider any $m \times n$ matrix B with $\text{rank}(B) \leq p$. Clearly, its null space has dimension no smaller than $n - p$. Therefore, we can find an orthonormal set $\{\mathbf{x}_1, \cdots, \mathbf{x}_{n-p}\}$ in the null space of B. Prove that for such vectors, we have

$$\|A - B\|_F^2 \quad \geq \quad \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2. \tag{2}$$

(c) (This requires slightly more subtle manipulations.) For any matrix A of dimension $m \times n$ and any *orthonormal* set of $n - p$ vectors in $\mathbb{C}^n$, denoted by $\{\mathbf{x}_1, \cdots, \mathbf{x}_{n-p}\}$, prove that

$$\sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \quad \geq \quad \sum_{j=p+1}^{r} \sigma_j^2. \tag{3}$$

Hint: Consider the case $m \geq n$ and the set of vectors $\{\mathbf{z}_1, \cdots, \mathbf{z}_{n-p}\}$, where $\mathbf{z}_k = V^H \mathbf{x}_k$. Express your formulas in terms of these and the SVD representation $A = U\Sigma V^H$.

(d) Briefly explain how *(a)-(c)* imply the desired statement.

Solution

(a) Let us collect the vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ (as columns) into an $n \times n$ matrix X. With this, we can express

$$\sum_{k=1}^{n} \|A\mathbf{x}_k\|^2 \quad = \quad \|AX\|_F^2. \tag{4}$$

Using the result that $\|A\|_F^2 = \text{trace}(A^H A)$, we find

$$\|AX\|_F^2 \quad = \quad \text{trace}((AX)^H AX) = \text{trace}(X^H A^H AX) = \text{trace}(A^H AXX^H), \tag{5}$$

where the last step is the property that $\text{trace}(AB) = \text{trace}(BA)$. But since by construction, X is a unitary matrix, we have that XX^H is simply the identity matrix. Hence, $\text{trace}(A^H AXX^H) = \text{trace}(A^H A)$, which completes the proof.

(b) Let us first expand our orthonormal set $\{\mathbf{x}_1, \cdots, \mathbf{x}_{n-p}\}$ to a full basis for $\mathbb{C}^n$ by including orthonormal vectors $\{\mathbf{x}_{n-p+1}, \cdots, \mathbf{x}_n\}$. Then, from Part *(a)*, we have

$$\|A - B\|_F^2 \quad = \quad \sum_{k=1}^{n} \|(A-B)\mathbf{x}_k\|^2 \geq \sum_{k=1}^{n-p} \|(A-B)\mathbf{x}_k\|^2, \tag{6}$$

where the last step is simply because all terms in the sum are non-negative. But by construction, $\{\mathbf{x}_1, \cdots, \mathbf{x}_{n-p}\}$ are in the null space of B, thus for them, $B\mathbf{x}_k = \mathbf{0}$, which implies $(A-B)\mathbf{x}_k = A\mathbf{x}_k$. This completes the proof.

(c) The first point of this exercise was to recall the often surprisingly useful "trick" that $\|\mathbf{y}\|^2 = \text{trace}(\mathbf{y}^H \mathbf{y})$, where of course the trace-operator does not do anything (yet). Applying this, we can express:

$$\begin{aligned} \|A\mathbf{x}_k\|^2 \quad &= \quad \text{trace}(\mathbf{x}_k^H A^H A \mathbf{x}_k) = \text{trace}(\mathbf{x}_k^H V \Sigma^H U^H U \Sigma V^H \mathbf{x}_k) && (7) \\ &= \quad \text{trace}(\mathbf{z}_k^H \Sigma^H \Sigma \mathbf{z}_k) = \text{trace}(\Sigma^H \Sigma \mathbf{z}_k \mathbf{z}_k^H), && (8) \end{aligned}$$

where in the last step, we have used the property $\text{trace}(AB) = \text{trace}(BA)$. Hence,

$$\begin{aligned} \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \quad &= \quad \sum_{k=1}^{n-p} \text{trace}(\Sigma^H \Sigma \mathbf{z}_k \mathbf{z}_k^H) && (9) \\ &= \quad \text{trace}(\Sigma^H \Sigma \sum_{k=1}^{n-p} \mathbf{z}_k \mathbf{z}_k^H) && (10) \\ &= \quad \sum_{k=1}^{n} \sigma_k^2 G_{kk}, && (11) \end{aligned}$$

where the last step is due to the fact that $\Sigma^H \Sigma$ is a *diagonal* matrix with entries σ_k^2, and where G_{ij} denote the entries of the matrix $G = \sum_{k=1}^{n-p} \mathbf{z}_k \mathbf{z}_k^H$. The matrix G is a projection matrix. As we have seen in an earlier homework problem, its trace is $n-p$ and its diagonal entries are non-negative and no larger than one. Under these constraints, it should be clear that the last expression is minimized if we select $G_{11} = G_{22} = \ldots = G_{pp} = 0$ and $G_{p+1,p+1} = \ldots = G_{nn} = 1$. Hence,

$$\begin{aligned} \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \quad &= \quad \sum_{k=1}^{n} \sigma_k^2 G_{kk} && (12) \\ &\geq \quad \sum_{k=p+1}^{n} \sigma_k^2 && (13) \\ &\geq \quad \sum_{k=p+1}^{r} \sigma_k^2, && (14) \end{aligned}$$

which completes the proof.

(d) Combining Parts *(b)* and *(c)*:

$$\|A - B\|_F^2 \quad \geq \quad \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \geq \sum_{j=p+1}^{r} \sigma_j^2 \tag{15}$$

shows that for *any* matrix B of rank p, we have the above lower bound. This is precisely the statement needed to complete the proof of the Eckart-Young theorem for the Frobenius norm.

Additional remark: Another proof of the Eckart-Young theorem (which works both for the Frobenius and the spectral norm) leverages the Weyl theorem, which states that for any two matrices C and D of the same dimension ($m \times n$, and assume w.l.o.g. $m \geq n$), we have that

$$\sigma_{i+j-1}(C+D) \quad \leq \quad \sigma_i(C) + \sigma_j(D), \quad \text{for } 1 \leq i,j \leq n, \text{ and } i+j-1 \leq n. \tag{16}$$

I am not aware of a simple proof of this theorem (the standard proof uses the variational characterization of eigenvalues). But suppose that B is of rank no larger than k, meaning that $\sigma_i(B) = 0$ for $i > k$. Then, setting $C = A - B$ and $D = B$, Weyl's theorem says that

$$\sigma_{i+k}(A) \quad \leq \quad \sigma_i(A-B) \quad \text{for } 1 \leq i \leq n-k, \tag{17}$$

and thus,

$$\|A-B\|_F^2 \quad \geq \quad \sum_{i=1}^{n-k} \sigma_i^2(A-B) \geq \sum_{i=k+1}^{n} \sigma_i^2(A). \tag{18}$$

Problem 2: Fourier (Review problem in view of our discussion of wavelets)

Suppose that a signal $x(t)$ satisfies

$$\int_{-\infty}^{\infty} x(t-n)x^*(t-m)dt \quad = \quad \begin{cases} 1, & \text{if } n = m \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

(In other words, the set of functions $\{x(t-n)\}_{n\in\mathbb{Z}}$ is an orthonormal set.) Show that then, its Fourier transform $X(\omega)$ must satisfy

$$\sum_{k\in\mathbb{Z}} |X(\omega + 2\pi k)|^2 \quad = \quad 1. \tag{20}$$

Solution

I am assuming that you have all come across this rather important property. For example, in Digital Communications, it is called the *Nyquist criterion,* at EPFL taught in the class *Principles of Digital Communications.*

Proof version 1:

An elegant approach to this problem is via the classical sampling theorem. To this end, we define the function $y(t) = \int_{-\infty}^{\infty} x(\tau+t)x^*(\tau)d\tau$. Then, we have that

$$\int_{-\infty}^{\infty} x(t-n)x^*(t-m)dt \quad = \quad y(m-n), \tag{21}$$

and thus, our assumption about the signal $x(t)$ is tantamout to requiring that the unit-interval samples of the signal $y(t)$ all vanish *except* the sample at time 0, which should be one.

To continue, from elementary Fourier properties, we know that[1]

$$Y(\omega) \quad = \quad |X(\omega)|^2. \tag{22}$$

[1]To see this, you may define $\tilde{x}(\tau) = x(-\tau)$ and thus write $y(t) = \int_{-\infty}^{\infty} x(\tau+t)\tilde{x}^*(-\tau)d\tau$. Doing a change of variable $\tau' = -\tau$ reveals that this integral is simply the convolution of $x(t)$ with $\tilde{x}^*(t)$. However, the Fourier transform of $\tilde{x}^*(t)$ is precisely $X^*(\omega)$. Using the convolution theorem of the Fourier transform gives the result.

Now, we invoke the standard sampling theorem. If we sample the signal $y(t)$ at unit time intervals, then the spectrum of the sampled signal is obtained by adding shifted copies of the spectrum $Y(\omega)$, where the shift corresponds to the sampling frequency. Since we are sampling at unit intervals, the sampling frequency is 2π, hence the spectral replica are placed at all multiples of 2π. That is, the spectrum of the sampled version of $y(t)$, let us call it $y_s(t)$, is simply given by

$$Y_s(\omega) \quad = \quad \sum_{k\in\mathbb{Z}} Y(\omega+2\pi k) = \sum_{k\in\mathbb{Z}} |X(\omega+2\pi k)|^2. \tag{23}$$

Clearly, the requirement that $\int_{-\infty}^{\infty} x(t-n)x^*(t-m)dt = \delta[n-m]$ means that $y(m-n) = \delta[n-m]$, or that all samples of $y(t)$ be zero, except the sample at zero itself. In other words, we want $y_s(t)$ to be a delta function (centered at $t=0$). The Fourier transform of a delta function is simply the constant 1 (for all frequencies ω). Hence, we require $Y_s(\omega) = 1$, which completes the proof.

Proof version 2:

Alternatively, we can also solve the problem from first principles, as follows: Since $X(w+2\pi k) = \int_{-\infty}^{\infty} x(t)e^{-j(w+2\pi k)t}dt$ and $X^*(w+2\pi k) = \int_{-\infty}^{\infty} x^*(t)e^{j(w+2\pi k)t}dt$, we have

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} |X(\omega+2\pi k)|^2 &= \sum_{k\in\mathbb{Z}} X(w+2\pi k)X^*(w+2\pi k) &(24)\\
&= \sum_{k\in\mathbb{Z}} \int_{-\infty}^{\infty} x(t_1)e^{-j(w+2\pi k)t_1}dt_1 \int_{-\infty}^{\infty} x^*(t_2)e^{j(w+2\pi k)t_2}dt_2 &(25)\\
&= \sum_{k\in\mathbb{Z}} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(t_1)x^*(t_2)e^{-j(w+2\pi k)(t_1-t_2)}dt_1dt_2 &(26)\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(t_1)x^*(t_2)e^{-jw(t_1-t_2)} \sum_{k\in\mathbb{Z}} e^{-j(t_1-t_2)2\pi k}dt_1dt_2 &(27)
\end{aligned}$$

Now, the key step is to calculate the sum inside the integral.

You have encountered this sum in the discussion of the *discrete-time Fourier transform (DTFT)*. (In the present class, we are not discussing this version of the Fourier transform.) This transform is defined as $Y(e^{j\omega}) = \sum_{n\in\mathbb{Z}} y[n]e^{-j\omega n}$. Now, let $y[k] = 1$, $\forall k \in \mathbb{Z}$. Then the Fourier transform of $y[k]$ is known to be

$$Y(e^{j\omega}) = \sum_{k=-\infty}^{\infty} y[k]e^{-j\omega k} = \sum_{k=-\infty}^{\infty} e^{-j\omega k} = 2\pi \sum_{\ell=-\infty}^{\infty} \delta(\omega - 2\pi\ell), \tag{28}$$

which is actually easier to see as an *inverse DTFT*, which is known to be $y[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} Y(e^{j\omega})e^{j\omega n}d\omega$, namely:

$$y[n] \quad = \quad \int_{-\pi}^{\pi} \left(\sum_{\ell=-\infty}^{\infty} \delta(\omega - 2\pi\ell) \right) e^{j\omega n}d\omega = \int_{-\pi}^{\pi} \delta(\omega)e^{j\omega n}d\omega = 1, \text{ for all } n. \tag{29}$$

Using this, we thus find that $\sum_{k\in\mathbb{Z}} e^{-j(t_1-t_2)2\pi k} = Y(e^{j((t_1-t_2)2\pi)}) = 2\pi\sum_{\ell=-\infty}^{\infty} \delta(2\pi(t_1-t_2-\ell))$.

But of course, this special (doubly) infinite sum appears in many other places, too.

Thus, we find

$$\begin{aligned}
\sum_{k\in\mathbb{Z}} |X(\omega+2\pi k)|^2 &= 2\pi \sum_{l=-\infty}^{\infty} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(t_1)x^*(t_2)e^{-jw(t_1-t_2)}\delta(2\pi(t_1-t_2-l))dt_1dt_2 &(30)\\
&= \sum_{l=-\infty}^{\infty} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x(t_1)x^*(t_2)e^{-jw(t_1-t_2)}\delta(2\pi(t_1-t_2-l))d(2\pi t_1)dt_2 &(31)
\end{aligned}$$

Note that only when $t_1 = t_2 + l$, the inner integral is none zero.

$$\sum_{k\in\mathbb{Z}} |X(\omega + 2\pi k)|^2 = \sum_{l=-\infty}^{\infty} \int_{-\infty}^{\infty} x(t_2 + l)x^*(t_2)e^{-jwl}dt_2 \quad (32)$$

$$= \sum_{l=-\infty}^{\infty} \int_{-\infty}^{\infty} x(t_2 + l)x^*(t_2)dt_2 e^{-jwl} \quad (33)$$

$$= 1 \quad (34)$$

In the last step, the integral is none zero only when $l = 0$.

Problem 3: Hilbert Space Projection Theorem

Given a Hilbert space H and a Hilbert subspace G, the Hilbert space projection theorem states that for every $x \in H$, there exists a unique $y \in G$ such that

(i) $x - y \in G^\perp$

(ii) $\|x - y\| = \inf_{u\in G} \|x - u\|$

Recall that $G^\perp = \{z \in H : \langle z, x\rangle = 0 \text{ for all } x \in G\}$.

Just like in class, prove that if y is indeed the minimizer of $\|x-u\|$ over all $u \in G$, then it must be true that $x - y \in G^\perp$, — except this time, justify every step as you "unpack" the norms into inner products, and use the properties of the inner product.

Solution

We want to show the orthogonality principle in an arbitrary Hilbert space, i.e., *only* exploiting the basic properties of a Hilbert space. Specifically, we want to prove the statement that

$$\|x - y\| = \inf_{u\in G} \|x - u\| \implies \langle x - y, z\rangle = 0, \text{ for all } z \in G. \quad (35)$$

Now, construct

$$y + \lambda z, \quad (36)$$

where $\lambda \in \mathbb{R}$, and note that $y + \lambda z$ is also in G. The trick is to decompose

$$\|x - (y + \lambda z)\|^2 = \langle x - (y + \lambda z), x - (y + \lambda z)\rangle \quad (37)$$

which is simply the definition of the symbol $\|\cdot\|^2$. Now, by the linearity of the inner product in the first argument, we can express this as

$$\|x - (y + \lambda z)\|^2 = \langle x - y, x - (y + \lambda z)\rangle - \lambda\langle z, x - (y + \lambda z)\rangle \quad (38)$$

$$= \langle x - (y + \lambda z), x - y\rangle^* - \lambda\langle x - (y + \lambda z), z\rangle^*, \quad (39)$$

where the second step follows from the property $\langle a, b\rangle = \langle b, a\rangle^*$. Now, using again the linearity of the inner product in the first argument, we further obtain

$$\|x - (y + \lambda z)\|^2 = \langle x - y, x - y\rangle^* - \lambda\langle z, x - y\rangle^* - \lambda\langle x - y, z\rangle^* + \langle -\lambda z, -\lambda z\rangle^*. \quad (40)$$

Again swapping arguments, and using the definition of $\|\cdot\|^2$, we find

$$\|x - (y + \lambda z)\|^2 = \|x - y\|^2 + \lambda^2\|z\|^2 - \lambda\langle x - y, z\rangle - \lambda\langle x - y, z\rangle^* \quad (41)$$

$$= \|x - y\|^2 + \lambda^2\|z\|^2 - \lambda Re\{\langle x - y, z\rangle\} \quad (42)$$

Now, we can use our assumption that y is chosen such that

$$\|x - y\| = \inf_{u \in G} \|x - u\|. \tag{43}$$

This can only be true if

$$\lambda^2 \|z\|^2 - \lambda Re\{\langle x - y, z\rangle\} \quad \geq \quad 0 \tag{44}$$

for *any* choice of λ. This is only possible if $Re\{\langle x - y, z\rangle\} = 0$, which can be understood along the following lines: Suppose first that $Re\{\langle x - y, z\rangle\} > 0$. Then, we can simply select

$$\lambda \quad < \quad \frac{Re\{\langle x - y, z\rangle\}}{\|z\|^2}. \tag{45}$$

and we will make the LHS of Equation (44) negative. The same can be shown for the case $Re\{\langle x - y, z\rangle\} < 0$, with appropriate choice of λ. Hence, Equation (44) can only be satisfied for all λ if $Re\{\langle x - y, z\rangle\} = 0$.

Now, we still have to show that $Im\{\langle x - y, z\rangle\} = 0$. This works exactly along the same lines, except that we now construct

$$y + j\lambda z, \tag{46}$$

where $\lambda \in \mathbb{R}$, and note that $y + j\lambda z$ is also in G.

Problem 4: Dual Basis

In class, we have mostly discussed orthonormal bases. Let $\{\varphi_n\}_{n \in Z}$ be a basis for the Hilbert space H. Then, for any vector $x \in H$, we have

$$x \quad = \quad \sum_n \langle x, \varphi_n \rangle \varphi_n \tag{47}$$

Now, suppose that $\{\varphi'_n\}_{n \in Z}$ is also a basis for H, but it is *not* orthonormal. Show that if we can find a so-called *dual* basis $\{\varphi''_n\}_{n \in Z}$ satisfying $\langle \varphi'_n, \varphi''_m \rangle = \delta(n - m)$ then for any vector $x \in H$, we have

$$x \quad = \quad \sum_n \langle x, \varphi''_n \rangle \varphi'_n. \tag{48}$$

Solution

For notational purposes, let us define

$$y \quad = \quad \sum_n \langle x, \varphi''_n \rangle \varphi'_n. \tag{49}$$

We will now show that $x = y$, which means more explicitly that

$$\|x - y\| \quad = \quad 0. \tag{50}$$

Equivalently, since $\{\varphi''_n\}_{n \in Z}$ is a basis of H, it is sufficient to show that

$$\langle x - y, \varphi''_k \rangle \quad = \quad 0, \tag{51}$$

for all k. (Can you prove this?) But then, observe that

$$\langle y, \varphi''_k \rangle \quad = \quad \langle \sum_n \langle x, \varphi''_n \rangle \varphi'_n, \varphi''_k \rangle \tag{52}$$

$$= \quad \sum_n \langle x, \varphi''_n \rangle \langle \varphi'_n, \varphi''_k \rangle \tag{53}$$

$$= \quad \langle x, \varphi''_k \rangle, \tag{54}$$

which establishes the claim. A natural follow-up problem is to find the dual basis $\{\varphi''_n\}_{n\in Z}$ for a given (non-orthogonal) basis $\{\varphi'_n\}_{n\in Z}$. This can be accomplished by a Gram-Schmidt procedure.

Problem 5: Minimum-norm Solutions

In this problem, we consider an *underdetermined* system of linear equations, i.e., $A\mathbf{x} = \mathbf{b}$, where A is a "fat" matrix ($m < n$) and $\mathbf{b}$ is chosen such that a solution exists. As you know, in this case, there exist infinitely many solutions. Prove that the one solution $\mathbf{x}$ that has the minimum 2-norm can be expressed as

$$\mathbf{x}_{MN} \quad = \quad V\Sigma^{-1}U^H\mathbf{b}, \tag{55}$$

where, as usual, the SVD of $A = U\Sigma V^H$, and Σ^{-1} is the matrix Σ where all non-zero diagonal entries are inverted.

Solution Let the SVD of $A = U\Sigma V^H$. Hence, U and V are unitary matrices, i.e. $U^{-1} = U^H$ and $V^{-1} = V^H$. Any $\mathbf{x}$ that satisfies $A\mathbf{x} = \mathbf{b}$ should also satisfy $U\Sigma V^H\mathbf{x} = \mathbf{b}$. Since A is a fat matrix $(m < n)$, there does not exists left inverse of Σ. And the only the first m diagonal entries of Σ can be non-zeros.

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m & \dots & 0 \end{bmatrix} \tag{56}$$

Let V_A denote the first m rows of V and V_B denote the last $n-m$ rows of V. Since the last $n-m$ columns of Σ are all zeros, it does not matter what V_B is. Since the dimensions of each row vector of A is n, it is possible to add $n-m$ linearly independent row vectors to A. The new SVD can be

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} U_A \\ U_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix} \begin{bmatrix} V_A \\ V_B \end{bmatrix}^H \tag{57}$$

Let $\mathbf{b}_B = B\mathbf{x}$ and $\mathbf{b}_A = \mathbf{b}$, then

$$\begin{bmatrix} A \\ B \end{bmatrix} \mathbf{x} = \begin{bmatrix} U_A \\ U_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix} \begin{bmatrix} V_A \\ V_B \end{bmatrix}^H \mathbf{x} = \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} \tag{58}$$

Now we have

$$\mathbf{x} = \begin{bmatrix} V_A \\ V_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U_A \\ U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} \tag{59}$$

Therefore, the square of the 2-norm of $\mathbf{x}$ is

$$||\mathbf{x}||_2^2 = \mathbf{x}^H\mathbf{x} \quad = \quad \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix}^H \begin{bmatrix} U_A \\ U_B \end{bmatrix} \begin{bmatrix} \Sigma_A^H & 0 \\ 0 & \Sigma_B^H \end{bmatrix}^{-1} \begin{bmatrix} V_A \\ V_B \end{bmatrix}^H \begin{bmatrix} V_A \\ V_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U_A \\ U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} \tag{60}$$

$$= \quad ||\mathbf{b}_A||_2^2 + ||\mathbf{b}_B||_2^2 \tag{61}$$

Thus, the 2-norm of $\mathbf{x}$ achieves minimum $||\mathbf{b}_A||_2 = ||\mathbf{b}||_2$, when $||\mathbf{b}_B||_2 = 0$. Also, $||\mathbf{b}_B||_2 = 0$ requires that every entry of $\mathbf{b}_B$ is 0. Hence in such case,

$$\mathbf{x}_{MN} = \begin{bmatrix} V_A \\ V_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U_A \\ U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b}_A \\ 0 \end{bmatrix} = V_A\Sigma_A^{-1}U_A^H\mathbf{b}_A = V\Sigma^{-1}U^H\mathbf{b} \tag{62}$$

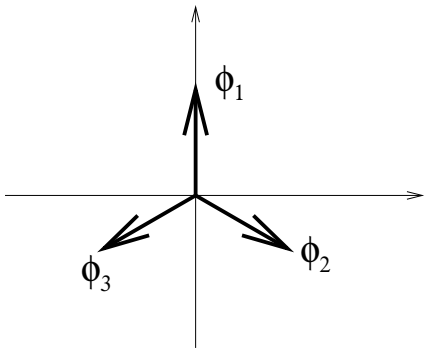

Figure 1: The three vectors $\phi_1, \phi_2,$ and ϕ_3 are at 120 degrees of each other and are of unit length each.

Problem 6: Frames

(a) We now turn to *overcomplete* expansions. The classic picture is given in Figure 1. In this picture, it is clear that every two-dimensional vector $\mathbf{x}$ can be written as

$$\mathbf{x} \quad = \quad a_1\phi_1 + a_2\phi_2 + a_3\phi_3 \tag{63}$$

in many different ways. Explicitly and for every two-dimensional vector $\mathbf{x}$, find the solution $\mathbf{a} = (a_1, a_2, a_3)^t$ with minimum energy,[2] i.e., minimizing $a_1^2 + a_2^2 + a_3^2$. Then, give a general formula for any finite-dimensional overcomplete expansion $\{\phi_n\}_{n=1}^N$ in $k-$dimensional space.

Solution

Rewriting the main equation, we find

$$\mathbf{x} \quad = \quad \Phi\mathbf{a}, \tag{64}$$

where the matrix Φ has the vectors ϕ_i as its columns. But this is simply an under-determined system of equations, just like in Problem 5. Hence, using the SVD $\Phi = U\Sigma V^H$, we can express

$$\mathbf{a}_{MN} \quad = \quad V\Sigma^{-1}U^H\mathbf{x}. \tag{65}$$

Assuming that all diagonal entries in Σ are strictly positive (which would be the case for any desirable frame!), this can be equivalently expressed via the so-called (Moore-Penrose) pseudo-inverse as:

$$\mathbf{a}_{MN} \quad = \quad \Phi^H(\Phi\Phi^H)^{-1}\mathbf{x}. \tag{66}$$

You can verify this easily by observing that $\Phi\Phi^H = U\Sigma^2U^H$, and thus, $(\Phi\Phi^H)^{-1} = U\Sigma^{-2}U^H$. Hence, $\Phi^H(\Phi\Phi^H)^{-1} = V\Sigma U^H U\Sigma^{-2}U^H = V\Sigma^{-1}U^H$. For the "Mercedes" frame, we can write the vectors as

$$\phi_1 \quad = \quad \binom{0}{1} \tag{67}$$

$$\phi_2 \quad = \quad \frac{1}{2}\binom{\sqrt{3}}{-1} \tag{68}$$

$$\phi_3 \quad = \quad \frac{1}{2}\binom{-\sqrt{3}}{-1} \tag{69}$$

(b) There is obviously no Parseval theorem, i.e., $x_1^2 + x_2^2 \neq a_1^2 + a_2^2 + a_3^2$. An overcomplete expansion is called a *frame* if there exist constants $0 < A \leq B < \infty$ such that

$$A\|x\|^2 \leq \sum_n |\langle x, \phi_n\rangle|^2 \leq B\|x\|^2. \tag{70}$$

[2]It should also be pointed out that in some Data Science applications, we don't want the minimum-energy solution, but the *sparsest* one, i.e., the one that has the fewest non-zero coefficients. In two dimensions, this is a trivial problem, but in N dimensions, there is no general simple solution, unfortunately...

Find the frame bounds A and B for the "Mercedes" frame above. *Note:* Because frames satisfy such a Parseval-like property, they are the most common overcomplete expansions. *Another note:* If $A = B$, the frame is called *tight*.

Solution For this frame, one really just has to plug in...

$$\begin{aligned} \sum_n |\langle x, \phi_n \rangle|^2 &= x_2^2 + \frac{1}{4}(3x_1^2 + x_2^2) + \frac{1}{4}(3x_1^2 + x_2^2) && (71) \\ &= \frac{3}{2}x_1^2 + \frac{3}{2}x_2^2 && (72) \\ &= \frac{3}{2}\|\mathbf{x}\|^2 && (73) \end{aligned}$$

So, we can select $A = B = \frac{3}{2}$, and conclude that the Mercedes-Benz frame is a tight frame.

Problem 7: Time–Frequency Representations

The elementary B-spline of degree 0 is the function $\beta^{(0)}(t) = 1$, for $-\frac{1}{2} \leq t < \frac{1}{2}$, and $\beta^{(0)}(t) = 0$ otherwise. The elementary B-spline of degree K is defined recursively as $\beta^{(K)} = \beta^{(K-1)} * \beta^{(0)}$. Find the Heisenberg box of the elementary B-splines of orders 0 and 1 (and 2, if you like). For each case, compare the size of the Heisenberg box to the lower bound (the uncertainty principle from class).

Solution For elementary B-splines of order 0, it can be verified that $||\beta^{(0)}||^2 = 1$, which implies the signal is normalized. Hence, the middle of the signal $\beta^{(0)}(t)$ is given by

$$m_t(\beta^{(0)}) = \int_{-\infty}^{\infty} t|\beta^{(0)}(t)|^2 dt = \int_{-1/2}^{1/2} t dt = 0 \tag{74}$$

And the squared time spread is given by

$$\sigma_t^2(\beta^{(0)}) = \int_{-\infty}^{\infty} (t - m_t(\beta^{(0)}))^2 |\beta^{(0)}(t)|^2 dt = \int_{-1/2}^{1/2} t^2 dt = \frac{1}{12} \tag{75}$$

Let $B^{(K)}$ denote the spectrum signal of $\beta^{(K)}$, for all degree K. Then the spectrum signal $B^{(0)}$ is the Fourier transform of $\beta^{(0)}$. Hence

$$B^{(0)} = \frac{2\sin(\omega/2)}{\omega} = \mathrm{sinc}(\omega/2) \tag{76}$$

The middle of the spectrum is given by

$$m_\omega(\beta^{(0)}) = \int_{-\infty}^{\infty} \omega \frac{1}{2\pi} |B^{(0)}(\omega)|^2 d\omega = 0 \tag{77}$$

And the squared frequency spread σ_ω^2 is given by

$$\sigma_\omega^2(\beta^{(0)}) = \int_{-\infty}^{\infty} (\omega - m_\omega(\beta^{(0)}))^2 \frac{1}{2\pi} |B^{(0)}(\omega)|^2 d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} \omega^2 \mathrm{sinc}^2(\omega/2) d\omega = \infty \tag{78}$$

For elementary B-splines of order 1, we have

$$\beta^{(1)} = \beta^{(0)} * \beta^{(0)} = \begin{cases} 1 + t, & t \in [-1, 0); \\ 1 - t, & t \in [0, 1); \\ 0, & \text{otherwise} \end{cases} \tag{79}$$

The squared norm of $\beta^{(1)}$ is

$$||\beta^{(1)}||^2 = \int_{-1}^{0}(1+t)^2dt + \int_{0}^{1}(1-t)^2dt = \frac{2}{3} \tag{80}$$

The middle of signal $\beta^{(1)}$ is given by

$$m_t(\beta^{(1)}) = \int_{-\infty}^{\infty} t|\beta^{(1)}(t)|^2dt = 0 \tag{81}$$

And the squared time spread is given by

$$\sigma_t^2(\beta^{(1)}) = \frac{1}{||\beta^{(1)}||^2}\int_{-\infty}^{\infty}(t - m_t(\beta^{(1)}))^2|\beta^{(1)}(t)|^2dt = \frac{1}{10} \tag{82}$$

The Fourier transform of $\beta^{(1)}$ is $B^{(1)} = \text{sinc}^2(\omega/2)$, which can be easily obtained by using convolution property of Fourier transform. The middle of the spectrum is given by

$$m_\omega(\beta^{(1)}) = \int_{-\infty}^{\infty}\omega\frac{1}{2\pi}|B^{(1)}(\omega)|^2d\omega = 0 \tag{83}$$

And the squared frequency spread σ_ω^2 is given by

$$\sigma_\omega^2(\beta^{(1)}) = \frac{1}{||\beta^{(1)}||^2}\int_{-\infty}^{\infty}(\omega - m_\omega(\beta^{(0)}))^2\frac{1}{2\pi}|B^{(1)}(\omega)|^2d\omega = \frac{1}{2\pi||\beta^{(1)}||^2}\int_{-\infty}^{\infty}\omega^2\text{sinc}^4(\omega/2)d\omega = 3 \tag{84}$$

The last integral can be looked up in any good integral table. However, for the Fourier cracks amongst you, it is an easy feat! As follows:

$$\begin{aligned}\frac{1}{2\pi}\int_{-\infty}^{\infty}\omega^2\text{sinc}^4(\omega/2)d\omega &= \frac{1}{2\pi}\int_{-\infty}^{\infty}|\underbrace{j\omega\text{sinc}^2(\omega/2)}_{=\Phi(\omega)}|^2d\omega &(85)\\ &= \int_{-\infty}^{\infty}|\phi(t)|^2d\omega &(86)\end{aligned}$$

by Parseval. But then, defining $\Psi(\omega) = \text{sinc}^2(\omega/2)$, we can observe that $\psi(t)$ is simply the triangle (the Fourier-inverse of the sinc-squared). Moreover, since $\Phi(\omega) = j\omega\Psi(\omega)$, from basic Fourier properties, we know that $\phi(t) = \frac{d}{dt}\psi(t)$. Hence, $\phi(t)$ is a simple piece-wise constant function:

$$\phi(t) = \begin{cases}1, & \text{for } -1 \le t < 0,\\ -1, & \text{for } 0 \le t < 1,\\ 0, & \text{otherwise.}\end{cases} \tag{87}$$

Hence, we find

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\omega^2\text{sinc}^4(\omega/2)d\omega = \int_{-\infty}^{\infty}|\phi(t)|^2d\omega = 2 \tag{88}$$

The size of Heisenberg Boxes for $\beta^{(0)}$ and $\beta^{(1)}$ are

$$\begin{aligned}\sigma_t(\beta^{(0)})\sigma_\omega(\beta^{(0)}) &= \frac{1}{\sqrt{12}}\times\sqrt{\infty} = \infty > \frac{1}{2} &(89)\\ \sigma_t(\beta^{(1)})\sigma_\omega(\beta^{(1)}) &= \frac{1}{\sqrt{10}}\times\sqrt{3} = \sqrt{\frac{3}{10}} > \frac{1}{2} &(90)\end{aligned}$$

Problem 8: Haar Wavelet

This problem is taken from Vetterli/Kovacevic, p. 295.

Consider the wavelet series expansion of continuous-time signals $f(t)$ and assume that $\psi(t)$ is the Haar wavelet.

(a) Give the expansion coefficients for $f(t) = 1, t \in [0,1]$, and 0 otherwise.

(b) Verify that for $f(t)$ as in Part *(a)*, $\sum_m \sum_n \|\langle \psi_{m,n}, f\rangle\|^2 = 1$ (i.e., Parseval's identity).

(c) Consider $f_1(t) = f(t - 2^{-i})$, where i is a positive integer. Give the range of scales over which expansion coefficients are non-zero. (Take $f(t)$ as in Part *(a)*.)

(d) Same as above, but now for $f_2(t) = f(t - 1/\sqrt{2})$. (Take $f(t)$ as in Part *(a)*.)

Solution

(a) The wavelet series expansion of a function $f(t)$ is given by

$$f(t) = \sum_{m,n} a_{m,n} \psi_{m,n}(t)$$

where $\psi_{m,n}(t) = 2^{-\frac{m}{2}} \psi(2^{-m}t - n)$ and $a_{m,n} = \langle f(t), \psi_{m,n}(t)\rangle$.
For the Haar expansion of $f(t) = \varphi(t)$ (the scaling function), the series coefficients are:

$$a_{m,n} = \int_0^1 \psi_{m,n}(t)\, dt = \begin{cases} 2^{-\frac{m}{2}} & m \geq 1 \quad n = 0 \\ 0 & \text{otherwise} \end{cases}$$

(b) $$\sum_m \sum_n |\langle \psi_{m,n}(t), f(t)\rangle|^2 = \sum_m \sum_n |a_{m,n}|^2 = \sum_{m=1}^{\infty} 2^{-m} = 1$$

(c) All coefficients of some scale m are equal to zero if and only if $2^{-i} = k\ 2^m$ for some integer k. Thus all coefficients are equal to zero for scales m such that $2^{-i-m} = k$ for some integer k. If $-i - m < 0$, the power 2^{-i-m} will not be an integer. Therefore, not all of the coefficients are zero for scales $m > -i$.

(d) All coefficients of some scale m are equal to zero if $\frac{1}{\sqrt{2}} = k\ 2^m$ for some integer k. Since there is no integer m for which this holds, there is no scale at which all of the coefficients are equal to zero.

Note that Parts *(c)* and *(d)* point to an undesirable feature of wavelets: For one and the same signal a particular shift (namely, of the form 2^{-i} for some integer i) leads to a very compact representation: many wavelet coefficients are zero. However, a different shift (like $\sqrt{2}$) does not have this nice property. People have tried to fix this by using efficient dedicated compression schemes for wavelet coefficients.