Problem Set 7
Signal Representation and Compression

For the Exercise Sessions on Dec 3 and 10 — **Due: Tue, Dec 16, 10am, on Moodle**

1 Problems for Class

Problem 1: Canonical Correlation Analysis

This is a follow-up to PCA. In PCA, we have n samples $\mathbf{x}^{(j)}$, for $j = 1, 2, \ldots, n$, where $\mathbf{x}^{(j)} \in \mathbb{R}^d$. Instead, here, we have n pairs of samples $(\mathbf{x}^{(j)}, \mathbf{y}^{(j)})$, for $j = 1, 2, \ldots, n$, where $\mathbf{x}^{(j)} \in \mathbb{R}^{d_x}$ and $\mathbf{y}^{(j)} \in \mathbb{R}^{d_y}$. For simplicity, we assume that the mean has already been subtracted, so that the samples satisfy $\sum_{j=1}^n \mathbf{x}^{(j)} = \mathbf{0}$ and $\sum_{j=1}^n \mathbf{y}^{(j)} = \mathbf{0}$. The goal of CCA is to extract a single projection from $\mathbf{x}^{(j)}$ and a single projection from $\mathbf{y}^{(j)}$ so that these two projections are as correlated as possible. That is, we seek to find those projection vectors $\mathbf{u}$ and $\mathbf{v}$ that maximize

$$\max_{\mathbf{u},\mathbf{v}} \frac{\sum_{j=1}^n (\mathbf{u}^T\mathbf{x}^{(j)})(\mathbf{v}^T\mathbf{y}^{(j)})}{\sqrt{\sum_{j=1}^n |\mathbf{u}^T\mathbf{x}^{(j)}|^2}\sqrt{\sum_{j=1}^n |\mathbf{v}^T\mathbf{y}^{(j)}|^2}}, \tag{1}$$

which is the classic ("Pearson") correlation coefficient. Show how we can find the optimizing choices of the vectors $\mathbf{u}$ and $\mathbf{v}$ from the empirical covariance matrices $R_{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^n (\mathbf{x}^{(j)})(\mathbf{x}^{(j)})^T$, $R_{\mathbf{Y}} = \frac{1}{n}\sum_{j=1}^n (\mathbf{y}^{(j)})(\mathbf{y}^{(j)})^T$, and $R_{\mathbf{XY}} = \frac{1}{n}\sum_{j=1}^n (\mathbf{x}^{(j)})(\mathbf{y}^{(j)})^T$.

Hint: First show that we can rewrite the objective as

$$\frac{\sum_{j=1}^n (\mathbf{u}^T\mathbf{x}^{(j)})(\mathbf{v}^T\mathbf{y}^{(j)})}{\sqrt{\sum_{j=1}^n |\mathbf{u}^T\mathbf{x}^{(j)}|^2}\sqrt{\sum_{j=1}^n |\mathbf{v}^T\mathbf{y}^{(j)}|^2}} = \frac{\mathbf{u}^T R_{\mathbf{XY}}\mathbf{v}}{\sqrt{\mathbf{u}^T R_{\mathbf{X}}\mathbf{u}}\sqrt{\mathbf{v}^T R_{\mathbf{Y}}\mathbf{v}}}.$$

Then, introduce a smart change of coordinates and apply the singular value decomposition.

Solution 1. We may write out:

$$\begin{aligned}\max_{u,v} \frac{\mathbb{E}[u^H XY^H v]}{\sqrt{\mathbb{E}[|u^H X|^2]}\sqrt{\mathbb{E}[|v^H Y|^2]}} &= \max_{u,v} \frac{u^H \mathbb{E}[XY^H]v}{\sqrt{u^H \mathbb{E}[XX^H]u}\sqrt{v^H \mathbb{E}[YY^H]v}} \\ &= \max_{u,v} \frac{u^H R_{XY} v}{\sqrt{u^H R_X u}\sqrt{v^H R_Y v}}\end{aligned}$$

Now, let $R_X^{1/2}$ be the matrix satisfying $R_X^{1/2}R_X^{1/2} = R_X$, and likewise, for $R_Y^{1/2}$. With this, define $\tilde{u} = R_X^{1/2}u$ and $\tilde{v} = R_Y^{1/2}v$. With this,

$$\begin{aligned}\max_{u,v} \frac{\mathbb{E}[u^H XY^H v]}{\sqrt{\mathbb{E}[|u^H X|^2]}\sqrt{\mathbb{E}[|v^H Y|^2]}} &= \max_{\tilde{u},\tilde{v}} \frac{\tilde{u}^H R_X^{-1/2} R_{XY} R_Y^{-1/2}\tilde{v}}{\sqrt{\tilde{u}^H\tilde{u}}\sqrt{\tilde{v}^H\tilde{v}}} \\ &= \max_{\|\tilde{u}\|=\|\tilde{v}\|=1} \tilde{u}^H R_X^{-1/2} R_{XY} R_Y^{-1/2}\tilde{v}. \end{aligned} \tag{2}$$

This last optimization problem is a classic question. Let us tackle this generally as

$$\max_{\|\tilde{u}\|=\|\tilde{v}\|=1} \tilde{u}^H A \tilde{v}.$$

We can solve this in a number of different ways. Closest to the class, let us simply leverage the singular value decomposition $A = U\Sigma V^H$. In terms of this, our problem becomes

$$\max_{\|\tilde{u}\|=\|\tilde{v}\|=1} \tilde{u}^H U\Sigma V^H \tilde{v}.$$

Now, define $\overline{u} = U^H\tilde{u}$ and $\overline{v} = V^H\tilde{v}$. Clearly, since U and V are unitary, the vectors $\overline{u}$ and $\overline{v}$ are also of unit length. Therefore, we may rewrite our problem as

$$\max_{\|\overline{u}\|=\|\overline{v}\|=1} \overline{u}^H \Sigma \overline{v}.$$

Without loss of generality, let us assume that the diagonal entries in Σ are ordered from largest singular value to smallest singular value. Then, the solution is easy: the optimal choices are $\overline{u}^* = (1, 0, \ldots, 0)^T$ and $\overline{v}^* = (1, 0, \ldots, 0)^T$. Of course, plugging back in, this means that the optimizing $\tilde{u}^*$ and $\tilde{v}^*$ are simply the left and right singular vectors corresponding to the largest singular value of the matrix A.

Now, to get back to our concrete problem, we have $A = R_X^{-1/2} R_{XY} R_Y^{-1/2}$. Let us denote the left and right singular vectors corresponding to the largest singular value of this matrix as $\tilde{u}_1(R_X^{-1/2} R_{XY} R_Y^{-1/2})$ and $\tilde{v}_1(R_X^{-1/2} R_{XY} R_Y^{-1/2})$. These are the optimizing choices for the optimization problem in Equation (2).

Finally, the optimizing choices for our original problem are therefore given by

$$\begin{aligned} u^* &= R_X^{-1/2} \tilde{u}_1(R_X^{-1/2} R_{XY} R_Y^{-1/2}) \\ v^* &= R_Y^{-1/2} \tilde{v}_1(R_X^{-1/2} R_{XY} R_Y^{-1/2}) \end{aligned} \tag{3}$$

Problem 2: Prediction and coding

After observing a binary sequence $u_1, \ldots, u_i$, that contains $n_0(u^i)$ zeros and $n_1(u^i)$ ones, we are asked to estimate the probability that the next observation, u_{i+1} will be 0. One class of estimators are of the form

$$\hat{P}_{U_{i+1}|U^i}(0|u^i) = \frac{n_0(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha} \quad \hat{P}_{U_{i+1}|U^i}(1|u^i) = \frac{n_1(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha}.$$

We will consider the case $\alpha = 1/2$, this is known as the Krichevsky–Trofimov estimator. Note that for $i = 0$ we get $\hat{P}_{U_1}(0) = \hat{P}_{U_1}(1) = 1/2$.

Consider now the joint distribution $\hat{P}(u^n)$ on $\{0,1\}^n$ induced by this estimator,

$$\hat{P}(u^n) = \prod_{i=1}^{n} \hat{P}_{U_i|U^{i-1}}(u_i|u^{i-1}).$$

(a) Show, by induction on n that, for any n and any $u^n \in \{0,1\}^n$,

$$\hat{P}(u_1, \ldots, u_n) \geq \frac{1}{2\sqrt{n}} \Big(\frac{n_0}{n}\Big)^{n_0} \Big(\frac{n_1}{n}\Big)^{n_1},$$

where $n_0 = n_0(u^n)$ and $n_1 = n_1(u^n)$.
[Hint: if $0 \leq m \leq n$, then $(1 + 1/n)^{n+1/2} \geq \frac{m+1}{m+1/2}(1 + 1/m)^m$]

(b) Conclude that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that

$$\operatorname{length} \mathcal{C}(u_1, \ldots, u_n) \leq n h_2\left(\frac{n_0(u^n)}{n}\right) + \frac{1}{2}\log n + 2,$$

with $h_2(x) = -x \log x - (1 - x)\log(1 - x)$.

(c) Show that if $U_1, \ldots, U_n$ are i.i.d. Bernoulli, then

$$\frac{1}{n}\mathbb{E}[\text{length}\, \mathcal{C}(U_1, \ldots, U_n)] \leq H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

Solution 2. (a) For $n = 1$, we have $\hat{P}(u_1) = \hat{P}_{U_1}(u_i) = \frac{1}{2}$. If $u_1 = 0$, $n_0(u_1) = 1$ and $n_1(u_1) = 0$. Hence, $\hat{P}(u_1) = \frac{1}{2} = \frac{1}{2\sqrt{n}}(\frac{n_0}{n})^{n_0}(\frac{n_1}{n})^{n_1}$. It is easy to show that for $u_1 = 1$, the inequality still holds with equality.

For $n = k \geq 1$, let's assume that $\hat{P}(u_1, \ldots, u_k) \geq \frac{1}{2\sqrt{k}}\left(\frac{n_0}{k}\right)^{n_0}\left(\frac{n_1}{k}\right)^{n_1}$. For $n = k+1$, it is sufficient to check $u_{k+1} = 0$, as the case $u_{i+1} = 1$ is the same if we also exchange the roles of n_0 and n_1. In this case, $n_0(u^{k+1}) = n_0(u^k) + 1$ and $n_1(u^{k+1}) = n_1(u^k)$.

$$\begin{aligned}
\hat{P}(u_1, \ldots, u_k, 0) &= \hat{P}_{U_{k+1}|U^k}(0|u^k)\hat{P}_{U^k}(u^k) \\
&\geq \frac{n_0(u^k) + \frac{1}{2}}{n_0(u^k) + n_1(u^k) + 1}\frac{1}{2\sqrt{k}}\left(\frac{n_0(u^k)}{k}\right)^{n_0(u^k)}\left(\frac{n_1(u^k)}{k}\right)^{n_1(u^k)} \\
&= \underbrace{\frac{(k+1)^{k+1/2}}{k^{k+1/2}}\frac{(n_0(u^k) + \frac{1}{2})n_0(u^k)^{n_0(u^k)}}{(n_0(u^k)+1)^{n_0(u^k)+1}}}_{f(u^k)}\frac{1}{2\sqrt{k+1}}\left(\frac{n_0(u^{k+1})}{k+1}\right)^{n_0(u^{k+1})}\left(\frac{n_1(u^{k+1})}{k+1}\right)^{n_1(u^{k+1})}
\end{aligned}$$

We need to show that $f(u^k) \geq 1$ for any $u^k \in \{0,1\}^k$, but this follows from the hint. Therefore, we proved that our induction hypothesis is true for any $n = k+1$, given the condition that $n = k$ cases is satisfied. By induction, we have for any integer $n \geq 1$

$$\hat{P}(u_1, \ldots, u_n) \geq \frac{1}{2\sqrt{n}}\left(\frac{n_0}{n}\right)^{n_0}\left(\frac{n_1}{n}\right)^{n_1},$$

Proof the hint: We need to show that:

$$\left(1 + \frac{1}{k}\right)^{k+1/2} \geq \underbrace{\frac{n_0(u^k)+1}{n_0(u^k)+\frac{1}{2}}\left(1 + \frac{1}{n_0(u^k)}\right)^{n_0(u^k)}}_{g(n_0(u^k)) = g(n_0)}.$$

Now, consider the function $g(x) = \frac{x+1}{x+\frac{1}{2}}(1+\frac{1}{x})^x$ for $x \geq 1$. Since we have that $n_0(u^k) \leq k$, if $g(x)$ is an increasing function then we would have:

$$\begin{aligned}
g(n_0(u^k)) \leq g(k) = \frac{k+1}{k+\frac{1}{2}}(1+\frac{1}{k})^k &= \frac{k+1}{(k+\frac{1}{2})\sqrt{1+\frac{1}{k}}}(1+\frac{1}{k})^{k+1/2} \\
&= \frac{\sqrt{k(k+1)}}{k+\frac{1}{2}}(1+\frac{1}{k})^{k+1/2} \\
&< \left(1+\frac{1}{k}\right)^{k+1/2},
\end{aligned}$$

and the result would follow (the last inequality is due to $\sqrt{k(k+1)} < \sqrt{k(k+1)+1/4} = k+1/2$). Hence, we just need to show that $g(x)$ is an increasing function, *i.e.* that $\frac{d}{dx}g(x) \geq 0$. A simple way of doing this is by showing that $\ln g(x)$ is an increasing function, which would then imply the result for $g(x)$. If we compute the differentiation of $\ln g(x)$, we get

$$\frac{d}{dx}\ln g(x) = \frac{1}{x+1} - \frac{1}{x+\frac{1}{2}} + \ln\left(1+\frac{1}{x}\right) - \frac{1}{x+1} = \ln(x+1) - \ln x - \frac{1}{x+\frac{1}{2}}$$

Now observe:

$$\ln(x+1) - \ln x = \int_x^{x+1} \frac{1}{u} du = \mathbb{E}\left[\frac{1}{U}\right],$$

where U is a unifom random variable between x and $x+1$. Also,

$$\frac{1}{x+1/2} = \frac{1}{\mathbb{E}[U]}.$$

Thus:

$$\frac{d}{dx} \ln g(x) = \mathbb{E}\left[\frac{1}{U}\right] - \frac{1}{\mathbb{E}[U]}$$

and the positivity of $\frac{d}{dx} \ln g(x)$ follows from the convexity of the function $u \to 1/u$ (and Jensen's inequality).

(b) Consider the code with length function $L(u^n) = \lceil -\log \hat{P}(u^n) \rceil$. We can check that such code satisfies the Kraft Inequity.

$$\sum_{u^n} 2^{-L(u^n)} = \sum_{u^n} 2^{-\lceil -\log \hat{P}(u^n) \rceil} \leq \sum_{u^n} \hat{P}(u^n) = 1$$

Hence, there exists a prefix-free code with length function $L(u^n)$.

$$\begin{aligned}
\text{length}\, \mathcal{C}(u_1, \ldots, u_n) = \lceil -\log \hat{P}(u^n) \rceil &\leq -\log \hat{P}(u^n) + 1 \\
&\leq -\log \left(\frac{1}{2\sqrt{n}} \left(\frac{n_0}{n}\right)^{n_0} \left(\frac{n_1}{n}\right)^{n_1} \right) + 1 \\
&= 2 + \frac{1}{2} \log n + n \left[-\frac{n_0}{n} \log(\frac{n_0}{n}) - \frac{n_1}{n} \log \frac{n_1}{n} \right] \\
&= 2 + \frac{1}{2} \log n + n h_2(\frac{n_0}{n})
\end{aligned}$$

(c) Let $\Pr(U_i = 0) = \theta$, $\forall i \in \{1, \ldots, n\}$. Since $U_1, \ldots, U_n$ are i.i.d, we have $\mathbb{E}[n_0(u^n)] = \sum_{i=1}^n \mathbb{E}[n_0(u_i)] = n\theta$ and $H(U_i) = h_2(\theta)$ for all i.

$$\begin{aligned}
\mathbb{E}[\text{length}\, \mathcal{C}(U_1, \ldots, U_n)] &\leq \mathbb{E}[n h_2(\frac{n_0(u^n)}{n}) + \frac{1}{2} \log n + 2] \\
&= n \mathbb{E}[h_2(\frac{n_0(u^n)}{n})] + \frac{1}{2} \log n + 2 \\
&\leq n h_2(\frac{\mathbb{E}[n_0(u^n)]}{n}) + \frac{1}{2} \log n + 2 \\
&= n h_2(\theta) + \frac{1}{2} \log n + 2 \\
&= n H(U_1) + \frac{1}{2} \log n + 2
\end{aligned}$$

Therefore,

$$\frac{1}{n} \mathbb{E}[\text{length}\, \mathcal{C}(U_1, \ldots, U_n)] \leq H(U_1) + \frac{1}{2n} \log n + \frac{2}{n}$$

2 The Homework

Problem 3: Inner Products

Consider the standard n-dimensional vector space $\mathbb{R}^n$.

1. Characterize the set of matrices W for which $\mathbf{y}^T W \mathbf{x}$ is a valid inner product for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

2. Prove that *every* inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ on $\mathbb{R}^n$ can be expressed as $\mathbf{y}^T W \mathbf{x}$ for an approriately chosen matrix W.

3. For a subspace of dimension $k < n$, spanned by the basis $\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k \in \mathbb{R}^n$, express the orthogonal projection operator (matrix) with respect to the general inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^T W \mathbf{x}$. *Hint:* For any vector $\mathbf{x} \in \mathbb{R}^n$, express its projection as $\widehat{\mathbf{x}} = \sum_{j=1}^k \alpha_j \mathbf{b}_j$.

Solution 3. 1. Looking at the lecture notes on the Hilbert space framework, an inner product must satisfy linearity properties, which clearly hold for all matrices W. The symmetry property $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ only holds if the matrix W is *symmetric,* i.e., $W^T = W$. The crucial requirement is the last property, namely, $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, with equality if and only if $\mathbf{x} = 0$. To tackle this, note that W has to be symmetric, so it has a spectral decomposition $W = U \Lambda U^H$. Hence, it is a clever idea to express the vectors $\mathbf{x}$ and $\mathbf{y}$ in terms of the eigenvectors of W. Then, clearly, if all eigenvalues of W are strictly positive, then the property is satisfied. Conversely, if there is a eigenvalue equal to zero, or a negative eigenvalue, then there exists a choice $\mathbf{x} \neq 0$ for which $\langle \mathbf{x}, \mathbf{x} \rangle = 0$. In conclusion, $\mathbf{y}^T W \mathbf{x}$ is a valid inner product if and only if W is a symmetric and positive definite.

2. To prove this, use the standard basis vectors to express $\mathbf{x} = x_1 \mathbf{e}_1 + \ldots + x_n \mathbf{e}_n$, and likewise for $\mathbf{y}$. Then, using the properties of the inner product, we find $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i,j} x_i y_i \langle \mathbf{e}_i, \mathbf{e}_j \rangle$. Notice that this is equal to

$$\sum_{i,j} x_i y_j \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \mathbf{x}^T \begin{bmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle & \langle \mathbf{e}_1, \mathbf{e}_2 \rangle & \ldots & \langle \mathbf{e}_1, \mathbf{e}_n \rangle \\ \langle \mathbf{e}_2, \mathbf{e}_1 \rangle & \langle \mathbf{e}_2, \mathbf{e}_2 \rangle & \ldots & \langle \mathbf{e}_2, \mathbf{e}_n \rangle \\ \vdots & \vdots & \vdots & \vdots \\ \langle \mathbf{e}_n, \mathbf{e}_1 \rangle & \langle \mathbf{e}_n, \mathbf{e}_2 \rangle & \ldots & \langle \mathbf{e}_n, \mathbf{e}_n \rangle \end{bmatrix} \mathbf{y}.$$

3. As we have seen in class, the error $\mathbf{x} - \widehat{\mathbf{x}}$ must be orthogonal to the estimate $\widehat{\mathbf{x}}$, or, equivalently, orthogonal to all of the basis vectors $\mathbf{b}_i$. That is,

$$\langle \mathbf{x} - \widehat{\mathbf{x}}, \mathbf{b}_i \rangle = 0. \tag{4}$$

Plugging in the hint $\widehat{\mathbf{x}} = \sum_{j=1}^k \alpha_j \mathbf{b}_j$, we get

$$\langle \mathbf{x} - \sum_{j=1}^k \alpha_j \mathbf{b}_j, \mathbf{b}_i \rangle = 0, \tag{5}$$

and using the standard properties of the inner product,

$$\langle \mathbf{x}, \mathbf{b}_i \rangle - \sum_{j=1}^k \alpha_j \langle \mathbf{b}_j, \mathbf{b}_i \rangle = 0. \tag{6}$$

Defining the $n \times k$ matrix

$$B = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k), \tag{7}$$

we can collect all k conditions (for $i = 1, 2, \ldots, k$) into

$$B^H W \mathbf{x} - B^H W B \alpha = 0, \tag{8}$$

where α denotes the column vector of all the coefficients α_i. Hence,

$$\alpha = \left(B^H W B \right)^{-1} B^H W \mathbf{x}, \tag{9}$$

where we note that $B^H W B$ is invertible since the vectors $\mathbf{b}_j$ constitute a basis. Finally, we observe that we can write

$$\widehat{\mathbf{x}} = B\alpha = B\left(B^H W B\right)^{-1} B^H W \mathbf{x}, \tag{10}$$

which is thus the desired projection matrix.

Problem 4: Johnson-Lindenstrauss for subgaussians

In this problem, we revisit the Johnson-Linderstrauss lemma.

(a) In preparation for this problem, establish the following facts:

- If U is a subexponential random variable with parameters (ν, b), then αU (where we assume $\alpha > 0$) is a subexponential random variable with parameters $(\alpha\nu, \alpha b)$.
- If U and V are independent subexponential random variables with parameters (ν_u, b_u) and (ν_v, b_v), respectively, then $U + V$ is a subexponential random variable with parameters $(\sqrt{\nu_u^2 + \nu_v^2}, \max(b_u, b_v))$.

In this problem, we reconsider the Johnson-Lindenstrauss Lemma (Lemma 10.5 in the lecture notes). The only change is that inside the real-valued $k \times d$ matrix X in the proof of the Lemma, we no longer assume that the entries are independent Gaussians. We still assume the entries X_{ij} to be independent. We also still assume that they each have mean zero and variance 1. But beyond this, we only assume that they are subgaussian with variance proxy σ^2.

To proceed, exactly as in the Johnson-Lindenstrauss Lemma, consider an arbitrary real-valued vector u of length d. As in the proof of the Johnson-Lindenstrauss Lemma, we define, for $i = 1, 2, \ldots, k$,

$$Z_i = \frac{1}{\|u\|_2} \sum_{j=1}^{d} u_j X_{ij}.$$

(b) Show the following facts (short justifications are sufficient, and you may refer freely to the lecture notes)

- The random variables Z_i are independent of each other.
- Each Z_i is subgaussian. Find the corresponding variance proxy.
- We have $\mathbb{E}[Z_i^2] = 1$.

To continue, we will need the following theorem:

Theorem. If Y is subgaussian with variance proxy σ^2, then Y^2 with mean $\mathbb{E}[Y^2]$ is subexponential with parameters $(c\sigma^2, d\sigma^2)$ for some absolute constants c and d.

(c) Exactly as in the proof of the Johnson-Lindenstrauss Lemma, we next need to analyze $S = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$. Leveraging the theorem, show that S is subexponential with mean 1 and find the corresponding parameters.

(d) Give a concentration bound, that is, an upper bound of the form

$$\mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^{k} Z_i^2 - 1\right| > \delta\right\} \leq \ldots$$

(e) Discuss the differences of the resulting lemma with respect to what is proved in the lecture notes.

Solution 4.

(a) We prove the two facts in turn, noting that the proof argument are identical to the proof of Parts (ii) and (iii) of Lemma 2.1 in the lecture notes:

- First, observe that the mean of αU is simply $\alpha\mu_u$, where μ_u denotes the mean of U. Hence, we need to study

$$\mathbb{E}[e^{\lambda(\alpha U - \alpha\mu_u)}] = \mathbb{E}[e^{\lambda\alpha(U-\mu_u)}].$$

  Now, since U is a subexponential random variable with parameters (ν, b), we know that we can upper bound this as

$$\mathbb{E}[e^{\lambda(\alpha U - \alpha\mu_u)}] = \mathbb{E}[e^{\lambda\alpha(U-\mu_u)}] \leq e^{(\lambda\alpha)^2\nu^2/2},$$

  as long as $|\lambda\alpha| < 1/b$. Now, we rewrite this just slightly. Namely, we can upper bound

$$\mathbb{E}[e^{\lambda(\alpha U - \alpha\mu_u)}] \leq e^{\lambda^2(\alpha\nu)^2/2},$$

  as long as $|\lambda| < 1/(\alpha b)$. Which is exactly the same as saying "αU is a subexponential random variable with parameters $(\alpha\nu, \alpha b)$".
- First, observe that the mean of $U + V$ is simply the sum of the means of U and V. Hence, looking at the definition of subexponential, we need to study

$$\mathbb{E}[e^{\lambda(U+V-\mu_u-\mu_v)}] = \mathbb{E}[e^{\lambda(U-\mu_u)}e^{\lambda(V-\mu_v)}] = \mathbb{E}[e^{\lambda(U-\mu_u)}]\mathbb{E}[e^{\lambda(V-\mu_v)}],$$

  where the last step follows because U and V are independent. Next, since U and V are subexponential, we can upper bound the two factors as

$$\mathbb{E}[e^{\lambda(U+V-\mu_u-\mu_v)}] = \mathbb{E}[e^{\lambda(U-\mu_u)}]\mathbb{E}[e^{\lambda(V-\mu_v)}] \leq e^{\nu_u^2\lambda^2/2}e^{\nu_v^2\lambda^2/2},$$

  which holds whenever $|\lambda| < 1/b_u$ and at the same time also $|\lambda| < 1/b_v$. Arranging terms, we can thus conclude that

$$\mathbb{E}[e^{\lambda(U+V-\mu_u-\mu_v)}] \leq e^{\left(\sqrt{\nu_u^2+\nu_v^2}\right)^2\lambda^2/2},$$

  whenever $|\lambda| < \min(1/b_u, 1/b_v)$. Which is exactly the same as saying "$U + V$ is a subexponential random variable with parameters $(\sqrt{\nu_u^2 + \nu_v^2}, \max(b_u, b_v))$".

(b) We take up the claims in turn:

- The random variables Z_i are independent of each other because Z_i are merely (weighted) sums of the X_{ij}, no X_{ij} appears in more than one of the Z_i, and the X_{ij} are by assumption independent of each other.
- Each Z_i is subgaussian simply because it is a (weighted) sum of subgaussian random variables, see Lemma 2.1 in the lecture notes. From that same lemma, we directly find that the variance proxy of Z_i is σ^2.
- We have $\mathbb{E}[Z_i^2] = \frac{1}{\|u\|_2^2}\sum_{j=1}^d u_j^2\mathbb{E}[X_{ij}^2]$, since the X_{ij} are independent of each other and have mean zero. Moreover, we have $\mathbb{E}[X_{ij}^2] = 1$ for all i and j, and thus, $\mathbb{E}[Z_i^2] = 1$.

(c) We know that each Z_i is subgaussian with variance proxy σ^2. Therefore, using the theorem, we know that each Z_i^2 is subexponential with mean 1 and parameters $(c\sigma^2, d\sigma^2)$. Moreover, all the Z_i^2 are independent of each other. Now, using the second half of Part (a), we can observe that $\sum_{i=1}^k Z_i^2$ is subexponential with mean k and parameters $(\sqrt{k}c\sigma^2, d\sigma^2)$. Then, using the first half

of Part (a), we can observe that $S = \frac{1}{k}\sum_{i=1}^{k} Z_i^2$ is subexponential with mean 1 and parameters $(\frac{c\sigma^2}{\sqrt{k}}, \frac{d\sigma^2}{k})$.

Alternatively, we could directly observe that the mean of S is 1 and write out, leveraging the fact that the Z_i^2 are independent of each other:

$$\begin{aligned}\mathbb{E}[e^{\lambda(S-1)}] &= \mathbb{E}[e^{\lambda(\frac{1}{k}\sum_{i=1}^{k} Z_i^2 - 1)}] \\ &= \prod_{i=1}^{k} \mathbb{E}[e^{\frac{\lambda}{k}(Z_i^2-1)}] \\ &\leq \left(e^{\frac{c^2}{2}\sigma^4\left(\frac{\lambda}{k}\right)^2}\right)^k = e^{\frac{c^2\sigma^4\lambda^2}{2k}},\end{aligned}$$

which holds for $|\frac{\lambda}{k}| < \frac{1}{d\sigma^2}$. That is, S with mean 1 is subexponential with parameters $(\frac{c\sigma^2}{\sqrt{k}}, \frac{d\sigma^2}{k})$.

(d) Here, we can directly leverage Lemma 2.6 from the Lecture Notes to conclude

$$\mathbb{P}\left\{\left|\frac{1}{k}\sum_{i=1}^{k} Z_i^2 - 1\right| > \delta\right\} \leq 2e^{-\frac{\delta^2 k}{2c^2\sigma^4}},$$

which holds whenever

$$\delta \leq \frac{\left(\frac{c\sigma^2}{\sqrt{k}}\right)^2}{\frac{d\sigma^2}{k}} = \frac{c^2}{d}\sigma^2.$$

More precisely, we apply Lemma 2.6 from the Lecture Notes separately to the positive and to the negative deviations. Since these are disjoint events, we can just add up the probabilities, which leads to the leading factor of 2 in our expression. This is an argument we have seen several times in the class.

(e) Discuss the differences of the resulting lemma with respect to what is proved in the lecture notes.

3 Additional Problems

Problem 5: Minimum-norm Solutions

In this problem, we consider an *underdetermined* system of linear equations, i.e., $A\mathbf{x} = \mathbf{b}$, where $A_{m\times n}$ is a "wide" 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

$$\mathbf{x}_{MN} \quad = \quad V\Sigma^{-1}U^H\mathbf{b}, \tag{11}$$

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.

Hint: Clearly, A is not a full-rank matrix, and thus cannot be inverted. However, it might be possible to *construct* a matrix A' such that $A'\mathbf{x} = \mathbf{b}'$ has a solution, A is a submatrix of A' and $\mathbf{b}$ is a subvector of $\mathbf{b}'$. What will be the norm of $\mathbf{x}$ in such a case?

Solution 5. Consider the singular value decomposition (SVD) $A = U\Sigma V^H$. Here, U is an $m \times m$ unitary matrix and Σ is given by

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_m \end{bmatrix}. \tag{12}$$

We assume that A is full rank, meaning that $\sigma_i > 0$ for all i. Finally, V is a matrix of dimension $n \times m$ whose columns are orthonormal, i.e., $V^H V = I_m$.

Note that $A\mathbf{x} = \mathbf{b}$ has infinitely many solutions. Our goal is to find the one solution with the smallest 2-norm.

We will now construct a new matrix which is of dimension $n \times n$ and of full rank. To this end, pick any unitary matrix U_B of dimension $(n-m) \times (n-m)$ and any diagonal matrix Σ_B, also of dimension $(n-m) \times (n-m)$, with strictly positive entries on the diagonal. Moreover, pick a matrix V_B of dimension $n \times (n-m)$ whose columns are orthonormal and at the same time orthogonal to all of the columns in V. (It is easy to see that such a matrix exists.) Define the matrix $B = U_B \Sigma_B V_B^H$. With this, we can now stack up A and B, and they satisfy

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} U & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma_B \end{bmatrix} \begin{bmatrix} V^H \\ V_B^H \end{bmatrix}, \tag{13}$$

where we also note that the last expression is indeed the singular value decomposition of the stacked matrix. Select any vector $\mathbf{b}_B$ of length $n-m$. Now consider solutions $\mathbf{x}$ to the system

$$\begin{bmatrix} A \\ B \end{bmatrix} \mathbf{x} = \begin{bmatrix} U & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma_B \end{bmatrix} \begin{bmatrix} V^H \\ V_B^H \end{bmatrix} \mathbf{x} = \begin{bmatrix} \mathbf{b} \\ \mathbf{b}_B \end{bmatrix}. \tag{14}$$

Since the matrix $\begin{bmatrix} A \\ B \end{bmatrix}$ is full rank by construction, thus invertible, the solution is given by

$$\mathbf{x} = \begin{bmatrix} V, V_B \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U & 0 \\ 0 & U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b} \\ \mathbf{b}_B \end{bmatrix}. \tag{15}$$

Moreover, note that this solution $\mathbf{x}$ also satisfies $A\mathbf{x} = \mathbf{b}$, so it is a solution to our original system of equations. The square of the 2-norm of $\mathbf{x}$ is

$$\begin{aligned} ||\mathbf{x}||_2^2 = \mathbf{x}^H \mathbf{x} &= \begin{bmatrix} \mathbf{b} \\ \mathbf{b}_B \end{bmatrix}^H \begin{bmatrix} U & 0 \\ 0 & U_B \end{bmatrix} \begin{bmatrix} \Sigma^H & 0 \\ 0 & \Sigma_B^H \end{bmatrix}^{-1} \begin{bmatrix} V^H \\ V_B^H \end{bmatrix} \begin{bmatrix} V, V_B \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U & 0 \\ 0 & U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b} \\ \mathbf{b}_B \end{bmatrix} \\ &= ||V\Sigma^{-1}U^H\mathbf{b}||_2^2 + ||V_B\Sigma_B^{-1}U_B^H\mathbf{b}_B||_2^2 \end{aligned} \tag{16}$$

Both summands in the last expression are non-negative. Since the first summand is fixed, the expression is minimized if we can make the second summand zero. To do so, we select $\mathbf{b}_B = \mathbf{0}$. Hence in such case,

$$\mathbf{x}_{MN} = \begin{bmatrix} V \\ V_B \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U & 0 \\ 0 & U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b} \\ \mathbf{0} \end{bmatrix} = V\Sigma^{-1}U^H\mathbf{b}. \tag{17}$$

Problem 6: 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

$$\|A\|_F^2 = \sum_{k=1}^{n} \|A\mathbf{x}_k\|^2. \tag{18}$$

(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

$$\|A - B\|_F^2 \quad \geq \quad \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2. \tag{19}$$

(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

$$\sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \quad \geq \quad \sum_{j=p+1}^{r} \sigma_j^2. \tag{20}$$

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 6. *(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

$$\sum_{k=1}^{n} \|A\mathbf{x}_k\|^2 \quad = \quad \|AX\|_F^2. \tag{21}$$

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

$$\|AX\|_F^2 \quad = \quad \text{trace}((AX)^H AX) = \text{trace}(X^H A^H AX) = \text{trace}(A^H AXX^H), \tag{22}$$

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

$$\|A - B\|_F^2 \quad = \quad \sum_{k=1}^{n} \|(A-B)\mathbf{x}_k\|^2 \geq \sum_{k=1}^{n-p} \|(A-B)\mathbf{x}_k\|^2, \tag{23}$$

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:

$$\begin{aligned} \|A\mathbf{x}_k\|^2 \quad &= \quad \text{trace}(\mathbf{x}_k^H A^H A\mathbf{x}_k) = \text{trace}(\mathbf{x}_k^H V\Sigma^H U^H U\Sigma V^H \mathbf{x}_k) && (24) \\ &= \quad \text{trace}(\mathbf{z}_k^H \Sigma^H \Sigma \mathbf{z}_k) = \text{trace}(\Sigma^H \Sigma \mathbf{z}_k \mathbf{z}_k^H), && (25) \end{aligned}$$

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

$$\begin{aligned} \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \quad &= \quad \sum_{k=1}^{n-p} \text{trace}(\Sigma^H \Sigma \mathbf{z}_k \mathbf{z}_k^H) && (26) \\ &= \quad \text{trace}(\Sigma^H \Sigma \sum_{k=1}^{n-p} \mathbf{z}_k \mathbf{z}_k^H) && (27) \\ &= \quad \sum_{k=1}^{n} \sigma_k^2 G_{kk}, && (28) \end{aligned}$$

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,

$$\begin{aligned} \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 &= \sum_{k=1}^{n} \sigma_k^2 G_{kk} && (29) \\ &\geq \sum_{k=p+1}^{n} \sigma_k^2 && (30) \\ &\geq \sum_{k=p+1}^{r} \sigma_k^2, && (31) \end{aligned}$$

which completes the proof.

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

$$\|A - B\|_F^2 \geq \sum_{k=1}^{n-p} \|A\mathbf{x}_k\|^2 \geq \sum_{j=p+1}^{r} \sigma_j^2 \tag{32}$$

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

$$\sigma_{i+j-1}(C+D) \leq \sigma_i(C) + \sigma_j(D), \quad \text{for } 1 \leq i,j \leq n, \text{ and } i+j-1 \leq n. \tag{33}$$

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

$$\sigma_{i+k}(A) \leq \sigma_i(A-B) \quad \text{for } 1 \leq i \leq n-k, \tag{34}$$

and thus,

$$\|A - B\|_F^2 \geq \sum_{i=1}^{n-k} \sigma_i^2(A-B) \geq \sum_{i=k+1}^{n} \sigma_i^2(A). \tag{35}$$

Problem 7: A Hilbert space of matrices

In this problem, we consider the set of matrices $A \in \mathbb{R}^{m \times n}$ with standard matrix addition and multiplication by scalar.

(a) Briefly argue that this is indeed a vector space, using the definition given in class.

(b) Show that $\langle A, B \rangle = \text{trace}(B^H A)$ is a valid inner product.

(c) Explicitly state the norm induced by this inner product. Is this a norm that you have encountered before?

(d) Consider as a further inner product candidate the form $\langle A, B \rangle = \text{trace}(B^H W A)$, where W is a square ($m \times m$) matrix. Give conditions on W such that this is a valid inner product. Explicit and detailed arguments are required for full credit.

Solution 7. *Note: In the following, we solve assuming the more general case of complex valued matrices. It should be easier to solve for the real-valued case.*

(a) We need to check some properties. Because the space is that of matrices,

1. Commutativity holds.
2. Associativity holds.
3. Distributivity holds.
4. The 0 element is the all 0's matrix $\mathbf{0} \in \mathbb{C}^{m\times n}$.
5. For all $A \in \mathbb{C}^{m\times n}$, we have that the element $-A \in \mathbb{C}^{m\times n}$ is such that $A + (-A) = \mathbf{0}$.
6. For all $A \in \mathbb{C}^{m\times n}$, we have that $I_{m\times m}A = A$.

So this is indeed a vector space.

(b) Here, we check the properties of an inner product space. Letting $A, B, C \in \mathbb{C}^{m\times n}, \alpha \in \mathbb{C}$, we have that

1. $\langle A + C, B\rangle = \mathrm{Tr}(B^H(A + C)) = \mathrm{Tr}(B^HA + B^HC) = \mathrm{Tr}(B^HA) + \mathrm{Tr}(B^HC) = \langle A, B\rangle + \langle C, B\rangle$, where we used the linearity of the trace operator.
2. $\langle \alpha A, B\rangle = \mathrm{Tr}(B^H\alpha A) = \alpha\mathrm{Tr}(B^HA) = \alpha\langle A, B\rangle$, where we used the linearity of the trace operator.
3. $\langle A, B\rangle = \mathrm{Tr}(B^HA) = \mathrm{Tr}((A^HB)^H) = \mathrm{Tr}(A^HB)^* = \langle B, A\rangle^*$, where we used the linearity of the trace operator and that conjugation is also linear.
4. We want $\langle A, A\rangle = \mathrm{Tr}(A^HA) \geq 0$. Since A^HA is normal, it is also positive semi-definite, and so all its eigenvalues are positive. One of the property of the trace is that it is equal to the sum of eigenvalues of the matrix considered. In our case, this means $\mathrm{Tr}(A^HA) \geq 0$ since the eigenvalues are all non-negative.

(c) We have that the norm is $\sqrt{\langle A, A\rangle} = \sqrt{\mathrm{Tr}(A^HA)} = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2}$, which we recognize to be the Frobenius norm. So $\sqrt{\langle A, A\rangle} = \|A\|_F$.

(d) We check that the properties of an inner product space hold, and add conditions on W when necessary.

1. $\langle A+C, B\rangle = \mathrm{Tr}(B^HW(A+C)) = \mathrm{Tr}(B^HWA+B^HWC) = \mathrm{Tr}(B^HA)+\mathrm{Tr}(B^HC) = \langle A, B\rangle+\langle C, B\rangle$, so no restriction on W necessary here.
2. $\langle \alpha A, B\rangle = \mathrm{Tr}(B^HW\alpha A) = \alpha\mathrm{Tr}(B^HWA) = \alpha\langle A, B\rangle$, so no restriction on W necessary here.
3. On one side we have $\langle A, B\rangle = \mathrm{Tr}(B^HWA)$. On the other side we have $\langle B, A\rangle^* = \mathrm{Tr}(A^HWB)^* = \mathrm{Tr}((A^HWB)^H) = \mathrm{Tr}(B^HW^HA)$. To have both sides equal, we need $W = W^H$, i.e., W should be Hermitian.
4. We want $\langle A, A\rangle = \mathrm{Tr}(A^HWA) \geq 0$. That is we would like A^HWA to be positive semi-definite. By definition, this would mean that for any $\mathbf{z} \in \mathbb{C}^n$, we want $\mathbf{z}^HA^HWA\mathbf{z} \geq 0$. Now note that $A\mathbf{z}$ is just another vector, so we can write that we want $\mathbf{z}^HA^HWA\mathbf{z} = (A\mathbf{z})^HW(A\mathbf{z}) \geq 0$ for all $\mathbf{z}$. So we conclude that this is the same as asking that W is positive semi-definite.

Hence, for the inner product $\langle A, B\rangle = \mathrm{Tr}(B^HWA)$ to be valid, we need W to be Hermitian and positive semi-definite.