Problem Set 7 **(Graded)** —*Due Tuesday, Dec 17, before class starts*

For the Exercise Sessions on Dec 3 and Dec 10

Last name	First name	SCIPER Nr	Points

Problem 1: 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 1. 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,

$$\langle \mathbf{x} - \widehat{\mathbf{x}}, \mathbf{b}_i \rangle = 0. \tag{1}$$

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

$$\langle \mathbf{x} - \sum_{j=1}^{k} \alpha_j \mathbf{b}_j, \mathbf{b}_i \rangle = 0, \tag{2}$$

and using the standard properties of the inner product,

$$\langle \mathbf{x}, \mathbf{b}_i \rangle - \sum_{j=1}^{k} \alpha_j \langle \mathbf{b}_j, \mathbf{b}_i \rangle = 0. \tag{3}$$

Defining the $n \times k$ matrix

$$B = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_k), \tag{4}$$

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

$$B^H W \mathbf{x} - B^H W B \alpha = 0, \tag{5}$$

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

$$\alpha = \left(B^H W B\right)^{-1} B^H W \mathbf{x}, \tag{6}$$

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

$$\widehat{\mathbf{x}} = B\alpha = B \left(B^H W B\right)^{-1} B^H W \mathbf{x}, \tag{7}$$

which is thus the desired projection matrix.

Problem 2: Canonical Correlation Analysis

Let $\mathbf{X}$ and $\mathbf{Y}$ be zero-mean real-valued random vectors with covariance matrices $R_{\mathbf{X}}$ and $R_{\mathbf{Y}}$, respectively. Moreover, let $R_{\mathbf{XY}} = \mathbb{E}[\mathbf{X}\mathbf{Y}^T]$. Our goal is to find vectors $\mathbf{u}$ and $\mathbf{v}$ such as to maximize the correlation between $\mathbf{u}^T\mathbf{X}$ and $\mathbf{v}^T\mathbf{Y}$, that is,

$$\max_{\mathbf{u},\mathbf{v}} \frac{\mathbb{E}[\mathbf{u}^T\mathbf{X}\mathbf{Y}^T\mathbf{v}]}{\sqrt{\mathbb{E}[|\mathbf{u}^T\mathbf{X}|^2]}\sqrt{\mathbb{E}[|\mathbf{v}^T\mathbf{Y}|^2]}}. \tag{8}$$

Show how we can find the optimizing choices of the vectors $\mathbf{u}$ and $\mathbf{v}$ from the problem parameters $R_{\mathbf{X}}, R_{\mathbf{Y}}$, and $R_{\mathbf{XY}}$.

Hint: Recall for the singular value decomposition that

$$\max_{\mathbf{v}} \frac{\|A\mathbf{v}\|}{\|\mathbf{v}\|} = \max_{\|\mathbf{v}\|=1} \|A\mathbf{v}\| = \sigma_1(A), \tag{9}$$

where $\sigma_1(A)$ denotes the maximum singular value of the matrix A. The corresponding maximizer is the right singular vector $\mathbf{v}_1$ (i.e., eigenvector of A^TA) corresponding to $\sigma_1(A)$.

Solution 2. Adapting the hint to this scenario, we prove that $\max_{\|u\|=\|v\|=1} u^H A v = \sigma_1(A)$ for a normal matrix A. Observe that $|u^H A v|^2 = (u^H A v)^H (u^H A v) = (u^H A v)(u^H A v)^H$, the first one gives $v^H A^H A v$ and the second one $u^H A A^H u$, both of these quantities are upper bounded by $\sigma_1(A)^2$ since this is the largest eigenvalue of $A^H A = A A^H$. Now finally if $u = v = v_1(A)$ then $u^H A v = \sigma_1(A)$ which achieves the bound.

We are now ready to prove the result using the decomposition $R_X = R_X^{1/2} R_X^{1/2}$

$$\begin{aligned}
\max_{u,v} \frac{\mathbb{E}[u^H X Y^H v]}{\sqrt{\mathbb{E}[|u^H X|^2]}\sqrt{\mathbb{E}[|v^H Y|^2]}} &= \max_{u,v} \frac{u^H \mathbb{E}[XY^H] v}{\sqrt{u^H \mathbb{E}[XX^H] u}\sqrt{v^H \mathbb{E}[YY^H] v}} \\
&= \max_{u,v} \frac{u^H R_{XY} v}{\sqrt{u^H R_X u}\sqrt{v^H R_Y v}} \\
&= \max_{u,v} \frac{u^H R_X^{-1/2} R_{XY} R_Y^{-1/2} v}{\sqrt{u^H u}\sqrt{v^H v}} \\
&= \max_{\|u\|=\|v\|=1} u^H R_X^{-1/2} R_{XY} R_Y^{-1/2} v \\
&= \sigma_1(R_X^{-1/2} R_{XY} R_Y^{-1/2})
\end{aligned} \tag{10}$$

and the corresponding values for original u and v (after equation (10) we apply the respective transformations $R_X^{-1/2}$ and $R_Y^{-1/2}$) are $u = R_X^{-1/2} v_1(R_X^{-1/2} R_{XY} R_Y^{-1/2})$ and $v = R_Y^{-1/2} v_1(R_X^{-1/2} R_{XY} R_Y^{-1/2})$.

Problem 3: 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 "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

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

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 3. 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.

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

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

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} U_A \\ U_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix} \begin{bmatrix} V_A \\ V_B \end{bmatrix}^H \tag{13}$$

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

$$\begin{bmatrix} A \\ B \end{bmatrix} \mathbf{x} = \begin{bmatrix} U_A \\ U_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix} \begin{bmatrix} V_A \\ V_B \end{bmatrix}^H \mathbf{x} = \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} \tag{14}$$

Now we have

$$\mathbf{x} = \begin{bmatrix} V_A \\ V_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U_A \\ U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} \tag{15}$$

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

$$\begin{aligned} ||\mathbf{x}||_2^2 = \mathbf{x}^H\mathbf{x} &= \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix}^H \begin{bmatrix} U_A \\ U_B \end{bmatrix} \begin{bmatrix} \Sigma_A^H & 0 \\ 0 & \Sigma_B^H \end{bmatrix}^{-1} \begin{bmatrix} V_A \\ V_B \end{bmatrix}^H \begin{bmatrix} V_A \\ V_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U_A \\ U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} && (16) \\ &= ||\mathbf{b}_A||_2^2 + ||\mathbf{b}_B||_2^2 && (17) \end{aligned}$$

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,

$$\mathbf{x}_{MN} = \begin{bmatrix} V_A \\ V_B \end{bmatrix} \begin{bmatrix} \Sigma_A & 0 \\ 0 & \Sigma_B \end{bmatrix}^{-1} \begin{bmatrix} U_A \\ U_B \end{bmatrix}^H \begin{bmatrix} \mathbf{b}_A \\ 0 \end{bmatrix} = V_A\Sigma_A^{-1}U_A^H\mathbf{b}_A = V\Sigma^{-1}U^H\mathbf{b} \tag{18}$$

Problem 4:

(Johnson-Lindenstrauss for subgaussians)

(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

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

(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.