ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
School of Computer and Communication Sciences

Learning Theory	Assignment date: August 19th, 2020, 08:15
Spring 2020	Due date: August 19th, 2020, 11:15

Final Exam – CS 526 – INM202

There are 4 general problems and 2 short questions. This exam is open-book (lecture notes, exercises, course materials) but no electronic devices allowed. Good luck!

Name: ____________________________

Section: ______________

Sciper No.: ______________

Problem 1	/ 22
Problem 2	/ 23
Problem 3	/ 25
Problem 4	/ 20
Problem 5	/ 10
Total	/100

Problem 1. *Tensors* (22 pts)

1. Consider the tensor $M = \begin{pmatrix}1\\2\end{pmatrix} \otimes \begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}1\\0\end{pmatrix} \otimes \begin{pmatrix}0\\1\end{pmatrix}$.

   (a) (2 pts) Write down the matrix components of M.

   (b) (5 pts) For the matrix M of part (a) exhibit an uncountable number of decompositions of the form $M = \vec{a} \otimes \vec{b} + \vec{c} \otimes \vec{d}$ using the rotation matrices

$$R = \begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}, \quad \theta \in \mathbb{R}.$$

2. (5 pts) Consider the following tensor decomposition

$$T = \begin{pmatrix}1\\1\\1\end{pmatrix} \otimes \begin{pmatrix}1\\0\\0\end{pmatrix} \otimes \begin{pmatrix}1\\0\\0\end{pmatrix} + \begin{pmatrix}0\\1\\2\end{pmatrix} \otimes \begin{pmatrix}0\\1\\0\end{pmatrix} \otimes \begin{pmatrix}0\\1\\1\end{pmatrix} + \begin{pmatrix}1\\3\\5\end{pmatrix} \otimes \begin{pmatrix}0\\1\\1\end{pmatrix} \otimes \begin{pmatrix}0\\1\\0\end{pmatrix}$$

   Is this decomposition unique? Justify your answer. What is the rank of T?

3. Let $\vec{a}_1, \vec{a}_2 \in \mathbb{R}^2$ be linearly independent and $\vec{b}_1, \vec{b}_2 \in \mathbb{R}^2$ be linearly independent as well. We define $T = \vec{a}_1 \otimes \vec{b}_1 \otimes \vec{c} + \vec{a}_2 \otimes \vec{b}_2 \otimes \vec{c}$ where $\vec{c} \in \mathbb{R}^2$ is not the zero vector.

   (a) (4 pts) Does Jennrich's theorem apply?

   (b) (6 pts) Prove that the tensor rank of T is 2.

Solution:

1. (a) $M = \begin{pmatrix}1 & 2\\ 2 & 2\end{pmatrix}$.

   (b) Using that $RR^T = I_2$:

$$M = \begin{pmatrix}1\\2\end{pmatrix} \otimes \begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}1\\0\end{pmatrix} \otimes \begin{pmatrix}0\\1\end{pmatrix} = \begin{pmatrix}1 & 1\\ 2 & 0\end{pmatrix}\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$$
$$= \begin{pmatrix}1 & 1\\ 2 & 0\end{pmatrix} R \cdot R^T \begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix} = \vec{a} \otimes \vec{b} + \vec{c} \otimes \vec{d}$$

   where

$$\begin{bmatrix}\vec{a} & \vec{c}\end{bmatrix} = \begin{pmatrix}1 & 1\\ 2 & 0\end{pmatrix} R = \begin{pmatrix}\cos\theta + \sin\theta & \cos\theta - \sin\theta\\ 2\cos\theta & -2\sin\theta\end{pmatrix};$$
$$\begin{bmatrix}\vec{b} & \vec{d}\end{bmatrix} = \begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix} R = \begin{pmatrix}\cos\theta & -\sin\theta\\ \cos\theta + \sin\theta & \cos\theta - \sin\theta\end{pmatrix}.$$

2. We have $T = \vec{a}_1 \otimes \vec{b}_1 \otimes \vec{c}_1 + \vec{a}_2 \otimes \vec{b}_2 \otimes \vec{c}_2 + \vec{a}_3 \otimes \vec{b}_3 \otimes \vec{c}_3$ where

$$\begin{bmatrix}\vec{a}_1 & \vec{a}_2 & \vec{a}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 1\\ 1 & 1 & 3\\ 1 & 2 & 5\end{bmatrix} \text{ has pairwise independent columns;}$$

$$\begin{bmatrix}\vec{b}_1 & \vec{b}_2 & \vec{b}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 1\\ 0 & 0 & 1\end{bmatrix} \text{ has linearly independent columns;}$$

$$\text{and } \begin{bmatrix}\vec{c}_1 & \vec{c}_2 & \vec{c}_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 1\\ 0 & 1 & 0\end{bmatrix} \text{ has linearly independent columns.}$$

By Jennrich's theorem the decomposition is therefore unique and the rank of T is 3.

3. (a) We have $T = \vec{a}_1 \otimes \vec{b}_1 \otimes \vec{c}_1 + \vec{a}_2 \otimes \vec{b}_2 \otimes \vec{c}_2$ where $\vec{c}_1 = \vec{c}_2 = \vec{c}$. We cannot invoke Jennrich's theorem because the vectors $\vec{c}_1, \vec{c}_2$ are not pairwise independent.

(b) The tensor rank is obviously less than or equal to 2. We will prove by contradiction that it cannot be equal to 1.

Assume the rank is one. Then there exist vectors $\vec{e}, \vec{f}, \vec{g}$ such that $T = \vec{e} \otimes \vec{f} \otimes \vec{g}$. Pick any vector $\vec{x}$ that is not orthogonal to $\vec{c}$. We have:

$$(\vec{e} \otimes \vec{f})(\vec{g}^T \vec{x}) = (\vec{a}_1 \otimes \vec{b}_1 + \vec{a}_2 \otimes \vec{b}_2)(\vec{c}^T \vec{x})$$

The matrix $(\vec{e} \otimes \vec{f})(\vec{g}^T \vec{x})$ has rank 0 or 1 while the matrix $(\vec{a}_1 \otimes \vec{b}_1 + \vec{a}_2 \otimes \vec{b}_2)(\vec{c}^T \vec{x})$ has rank 2 because $\vec{a}_1 \otimes \vec{b}_1 + \vec{a}_2 \otimes \vec{b}_2$ has rank 2 and $\vec{c}^T \vec{x} \neq 0$. This is a contradiction.

Problem 2. *Tensor decomposition & estimation of a sensing matrix* (23 pts)

Let $N \geq K$ two positive integers. Define the $N \times K$ real matrix $A := \begin{bmatrix} \vec{\mu}^{(1)} & \vec{\mu}^{(2)} & \cdots & \vec{\mu}^{(K)} \end{bmatrix}$ where the column vectors $\vec{\mu}^{(k)} = (\mu_\alpha^k)_{\alpha=1}^N$, $k = 1 \dots K$, are (fixed) N-dimensional linearly independent vectors.

Let $\vec{h} = (h_k)_{k=1}^K$ be a random vector whose components h_k's are independently (but not necessarily identically) distributed. We assume that $\forall k : \mathbb{E}[h_k] = \mathbb{E}[h_k^3] = 0$ and $\mathbb{E}[h_k^2], \mathbb{E}[h_k^4]$ are finite positive. We define the *excess kurtoses* $\mathcal{K}_k = \frac{\mathbb{E}[h_k^4]}{\mathbb{E}[h_k^2]^2} - 3$. If h_k has a zero-mean Gaussian distribution then $\mathcal{K}_k = 0$, so $\mathcal{K}_k$ can be essentially viewed as a measure of non-Gaussianity.

We are given L observations $\vec{y}^{(\ell)} = (y_\alpha^\ell)_{\alpha=1}^N := A\vec{h}^{(\ell)}$ where $\vec{h}^{(1)}, \vec{h}^{(2)}, \dots, \vec{h}^{(L)} \overset{\text{i.i.d.}}{\sim} \vec{h}$. Except for what is known on the distribution of $\vec{h}$, we don't know anything on the input vectors $\vec{h}^{(1)}, \vec{h}^{(2)}, \dots, \vec{h}^{(L)}$. The goal of the exercise is to show how to recover the columns of the sensing matrix A from these L observations.

1. (2 pts) Let $\vec{y} := A\vec{h}$. We define $\widehat{S}$ and $\widehat{F}$ the empirical estimates (using the L observations $\vec{y}^{(\ell)}$) of the second-moment matrix $S := \mathbb{E}[\vec{y} \otimes \vec{y}]$ and the fourth-moment tensor $F := \mathbb{E}[\vec{y} \otimes \vec{y} \otimes \vec{y} \otimes \vec{y}]$.

   Write down expressions for the components $\widehat{S}_{\alpha\beta}$ of $\widehat{S}$ and $\widehat{F}_{\alpha\beta\gamma\delta}$ of $\widehat{F}$ in terms of the components of $\vec{y}^{(1)}, \vec{y}^{(2)}, \dots, \vec{y}^{(L)}$.

2. (3 pts) From now on we suppose that $\widehat{S}$ and $\widehat{F}$ are good estimates of S and F, respectively. Prove the following identities:

$$S = \sum_{k=1}^K \mathbb{E}[h_k^2]\, \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \ ;$$

$$\begin{aligned} F = \sum_{k=1}^K \mathbb{E}[h_k^4]\, \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} + \sum_{1 \leq j \neq k \leq K} \mathbb{E}[h_j^2]\mathbb{E}[h_k^2] \Big(& \vec{\mu}^{(j)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \\ & + \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \\ & + \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \Big). \end{aligned}$$

3. (3 pts) We now form the tensor T with components

$$T_{\alpha\beta\gamma\delta} := F_{\alpha\beta\gamma\delta} - S_{\alpha\beta}S_{\gamma\delta} - S_{\alpha\gamma}S_{\beta\delta} - S_{\alpha\delta}S_{\beta\gamma} \ .$$

   Use the previous question to show that

$$T = \sum_{k=1}^K \mathcal{K}_k \mathbb{E}[h_k^2]^2\, \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \ .$$

4. (3 pts) Show that $S = UDU^T$ with $U = \begin{bmatrix} \vec{u}^{(1)} & \vec{u}^{(2)} & \cdots & \vec{u}^{(K)} \end{bmatrix} \in \mathbb{R}^{N\times K}$ a matrix with orthonormal columns, and $D = \mathrm{Diag}(d_1, d_2, \dots, d_K)$ a diagonal matrix with diagonal entries $d_1 \geq d_2 \geq \cdots \geq d_K > 0$.

5. (3 pts) Define the vectors $\vec{v}^{(k)} := \sqrt{\mathbb{E}[h_k^2]}\, W^T \vec{\mu}^{(k)}$, $k = 1 \dots K$, where $W = UD^{-\frac{1}{2}}$ and the tensor $\widetilde{T} := \sum_{k=1}^{K} \mathcal{K}_k\, \vec{v}^{(k)} \otimes \vec{v}^{(k)} \otimes \vec{v}^{(k)} \otimes \vec{v}^{(k)}$.

   Explain how to obtain the components $\widetilde{T}_{\alpha\beta\gamma\delta}$ of $\widetilde{T}$ from those of T, i.e., write down the formula relating them. How is this process (the transformation of T into $\widetilde{T}$) called?

6. (2 pts) As we have seen in class, the set of vectors $\{\vec{v}^{(1)}, \dots, \vec{v}^{(K)}\}$ is orthonormal and we can try to recover them using the tensor power method.

   What happens if one of the excess kurtosis $\mathcal{K}_k$ is zero?

7. (4 pts) From now on we suppose that all the excess kurtoses are nonzero.

   Write a small pseudo-code for the power method applied to $\widetilde{T}$ to recover $\mathcal{K}_k$ and (up to a plus or minus sign) $\vec{v}^{(k)}$ for $k = 1 \dots K$.

8. (3 pts) Assume that we also know the second moments $\mathbb{E}[h_k^2]$ for $k = 1 \dots K$.

   After having recovered $\mathcal{K}_k$ and $\pm\vec{v}^{(k)}$ for $k = 1 \dots K$ with the power method, how do you recover $\pm\vec{\mu}^{(k)}$ (so up to a plus or minus sign) for $k = 1 \dots K$?

Solution:

1. The empirical estimate $\widehat{S} = \frac{1}{L}\sum_{\ell=1}^{L} \vec{y}^{(\ell)} \otimes \vec{y}^{(\ell)}$ of S has components

$$\widehat{S}_{\alpha\beta} = \frac{1}{L}\sum_{\ell=1}^{L} y_\alpha^\ell y_\beta^\ell \,.$$

   The empirical estimate $\widehat{F} = \frac{1}{L}\sum_{\ell=1}^{L} \vec{y}^{(\ell)} \otimes \vec{y}^{(\ell)} \otimes \vec{y}^{(\ell)} \otimes \vec{y}^{(\ell)}$ of F has components

$$\widehat{F}_{\alpha\beta\gamma\delta} = \frac{1}{L}\sum_{\ell=1}^{L} y_\alpha^\ell y_\beta^\ell y_\gamma^\ell y_\delta^\ell \,.$$

2. Remember that $\vec{y} = \sum_{k=1}^{k} h_k \vec{\mu}^{(k)}$. By expanding the tensor products we get:

$$S = \mathbb{E}[\vec{y} \otimes \vec{y}] = \sum_{j=1}^{K}\sum_{k=1}^{K} \mathbb{E}[h_j h_k] \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \,; \tag{1}$$

$$F = \mathbb{E}[\vec{y} \otimes \vec{y} \otimes \vec{y} \otimes \vec{y}] = \sum_{i=1}^{K}\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{\ell=1}^{K} \mathbb{E}[h_i h_j h_k h_\ell] \vec{\mu}^{(i)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(\ell)} \,. \tag{2}$$

The components of h are independent and centered so $\mathbb{E}[h_j h_k] = \mathbb{E}[h_j]\mathbb{E}[h_k] = 0$ if $j \neq k$. Hence (1) simplifies:

$$S = \mathbb{E}[\vec{y} \otimes \vec{y}] = \sum_{k=1}^{K} \mathbb{E}[h_k^2] \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \; .$$

Similarly, if one of the indices i, j, k, ℓ is distinct of all the others then $\mathbb{E}[h_i h_j h_k h_\ell] = 0$. Therefore $\mathbb{E}[h_i h_j h_k h_\ell]$ is nonzero if, and only if, $i = j = k = \ell$ or $i = j \neq k = \ell$, $i = k \neq j = \ell$, $i = \ell \neq j = k$. Hence (2) reads:

$$\begin{aligned} F = \sum_{k=1}^{K} \mathbb{E}[h_k^4] \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} &+ \sum_{1 \leq i \neq k \leq K} \mathbb{E}[h_i^2]\mathbb{E}[h_k^2] \vec{\mu}^{(i)} \otimes \vec{\mu}^{(i)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \\ &+ \sum_{1 \leq k \neq j \leq K} \mathbb{E}[h_k^2]\mathbb{E}[h_j^2] \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \\ &+ \sum_{1 \leq i \neq k \leq K} \mathbb{E}[h_i^2]\mathbb{E}[h_k^2] \vec{\mu}^{(i)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(i)} \\ = \sum_{k=1}^{K} \mathbb{E}[h_k^4] \, \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} &+ \sum_{1 \leq j \neq k \leq K} \mathbb{E}[h_j^2]\mathbb{E}[h_k^2] \Big(\vec{\mu}^{(j)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \\ &\qquad + \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \\ &\qquad + \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \Big). \end{aligned}$$

3. We have seen that $F = \sum_{k=1}^{K} \mathbb{E}[h_k^4] \, \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} + E$ where

$$\begin{aligned} E := \sum_{1 \leq j \neq k \leq K} \mathbb{E}[h_j^2]\mathbb{E}[h_k^2] \Big(&\vec{\mu}^{(j)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} + \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \\ &+ \vec{\mu}^{(j)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(j)} \Big) \; . \end{aligned}$$

The components of E satisfy:

$$\begin{aligned} E_{\alpha\beta\gamma\delta} &= \sum_{1 \leq j \neq k \leq K} \mathbb{E}[h_j^2]\mathbb{E}[h_k^2] (\mu_\alpha^j \mu_\beta^j \mu_\gamma^k \mu_\delta^k + \mu_\alpha^j \mu_\beta^k \mu_\gamma^j \mu_\delta^k + \mu_\alpha^j \mu_\beta^k \mu_\gamma^k \mu_\delta^j) \\ &= \Big(\sum_{j=1}^{K} \mathbb{E}[h_j^2] \mu_\alpha^j \mu_\beta^j \Big) \Big(\sum_{k=1}^{K} \mathbb{E}[h_k^2] \mu_\gamma^k \mu_\delta^k \Big) + \Big(\sum_{j=1}^{K} \mathbb{E}[h_j^2] \mu_\alpha^j \mu_\gamma^j \Big) \Big(\sum_{k=1}^{K} \mathbb{E}[h_k^2] \mu_\beta^k \mu_\delta^k \Big) \\ &\quad + \Big(\sum_{j=1}^{K} \mathbb{E}[h_j^2] \mu_\alpha^j \mu_\delta^j \Big) \Big(\sum_{k=1}^{K} \mathbb{E}[h_k^2] \mu_\beta^k \mu_\gamma^k \Big) - 3 \sum_{k=1}^{K} \mathbb{E}[h_k^2]^2 \mu_\alpha^k \mu_\beta^k \mu_\gamma^k \mu_\delta^k \\ &= S_{\alpha\beta} S_{\gamma\delta} + S_{\alpha\gamma} S_{\beta\delta} + S_{\alpha\delta} S_{\beta\gamma} - 3 \Big(\sum_{k=1}^{K} \mathbb{E}[h_k^2]^2 \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \otimes \vec{\mu}^{(k)} \Big)_{\alpha\beta\gamma\delta} \; . \end{aligned}$$

It directly follows that

$$\begin{aligned}T_{\alpha\beta\gamma\delta} &= \left(\sum_{k=1}^{K}\mathbb{E}[h_k^4]\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\right)_{\alpha\beta\gamma\delta} \\ &\qquad\qquad - 3\left(\sum_{k=1}^{K}\mathbb{E}[h_k^2]^2\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\right)_{\alpha\beta\gamma\delta} \\ &= \left(\sum_{k=1}^{K}\mathcal{K}_k\mathbb{E}[h_k^2]^2\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\otimes\vec{\mu}^{(k)}\right)_{\alpha\beta\gamma\delta} .\end{aligned}$$

4. The matrix S is symmetric positive semidefinite so it can be diagonalised in an orthonormal basis: $S = \sum_{k=1}^{N} d_k\vec{u}^{(k)}\otimes\vec{u}^{(k)}$ with $\begin{bmatrix}\vec{u}^{(1)} & \vec{u}^{(2)} & \cdots & \vec{u}^{(N)}\end{bmatrix} \in \mathbb{R}^{N\times N}$ an orthonormal matrix and $d_1 \geq d_2 \geq \cdots \geq d_N \geq 0$. Besises S has rank K so exactly K of its eigenvalues are nonzero: $S = \sum_{k=1}^{K} d_k\vec{u}^{(k)}\otimes\vec{u}^{(k)} = UDU^T$ with $U = \begin{bmatrix}\vec{u}^{(1)} & \vec{u}^{(2)} & \cdots & \vec{u}^{(K)}\end{bmatrix} \in \mathbb{R}^{N\times K}$ and $D = \mathrm{Diag}(d_1, d_2, \dots, d_K)$.

5. By definition of the vectors $\vec{v}^{(k)}$ we have:

$$\widetilde{T} = \sum_{k=1}^{K}\mathcal{K}_k\mathbb{E}[h_k^2]^2(W^T\vec{\mu}^{(k)})\otimes(W^T\vec{\mu}^{(k)})\otimes(W^T\vec{\mu}^{(k)})\otimes(W^T\vec{\mu}^{(k)}) .$$

   So the components of $\widetilde{T}$ are given by the formula

$$\widetilde{T}_{\alpha\beta\gamma\delta} = \sum_{\alpha',\beta',\gamma',\delta'} W_{\alpha'\alpha}W_{\beta'\beta}W_{\gamma'\gamma}W_{\delta'\delta}\, T_{\alpha'\beta'\gamma'\delta'} .$$

   This transformation of T into $\widetilde{T}$ is called a whitening process.

6. If $\mathcal{K}_k$ is zero then it is impossible to recover $\vec{v}^{(k)}$ with the tensor power method and, as a consequence, to recover $\vec{\mu}^{(k)}$ the k^{th} column of A.

7. The pseudocode for the tensor power method is given in Algorithm 1.

8. By definition $\vec{v}^{\,(k)} = \sqrt{\mathbb{E}[h_k^2]}W^T\vec{\mu}^{(k)} = \sqrt{\mathbb{E}[h_k^2]}D^{-\frac{1}{2}}U^T\vec{\mu}^{(k)}$ so

$$UU^T\vec{\mu}^{(k)} = UD^{\frac{1}{2}}\frac{\vec{v}^{\,(k)}}{\sqrt{\mathbb{E}[h_k^2]}} .$$

   Finally, $\vec{\mu}^{(k)}$ belongs to the subspace spanned by the columns of U so $UU^T\vec{\mu}^{(k)} = \vec{\mu}^{(k)}$. We conclude that $\vec{\mu}^{(k)} = UD^{\frac{1}{2}}\vec{v}^{\,(k)}/\sqrt{\mathbb{E}[h_k^2]}$. Of course, we only know $\vec{v}^{\,(k)}$ up to a plus or minus sign and we will recover $\vec{\mu}^{(k)}$ up to a plus or minus sign too.

Algorithm 1 Tensor power method

```
procedure PowerMethod($\widetilde{T}, K, T_{\max}$)
    vectors $\leftarrow$ [ ]
    kurtoses $\leftarrow$ [ ]
    for $i = 1$ to $K$ do
        $v \leftarrow$ normallyDistributedVector(size = $K$)
        for $j = 1$ to $T_{\max}$ do                    ▷ $T_{\max}$ iterations of the power method.
            for $\alpha = 1$ to $K$ do
                $v[\alpha] \leftarrow \sum_{\beta,\gamma,\delta=1}^{K} \widetilde{T}_{\alpha,\beta,\gamma,\delta} v[\beta] v[\gamma] v[\delta]$
            end for
            $v \leftarrow {}^{v}/_{\|v\|}$                                  ▷ Renormalizing $v$
        end for
        $\mathcal{K} \leftarrow \sum_{,\alpha,\beta,\gamma,\delta=1}^{K} \widetilde{T}_{\alpha,\beta,\gamma,\delta} v[\alpha] v[\beta] v[\gamma] v[\delta]$
        kurtoses.append($\mathcal{K}$)
        vectors.append($v$)
        $\widetilde{T} \leftarrow \widetilde{T} - \mathcal{K}\, v^{\otimes 4}$          ▷ Subtracting the recovered rank-one tensor from $\widetilde{T}$
    end for
    return vectors, kurtoses
end procedure
```

Problem 3. *Stability implies Generalization* (25 pts) Let $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ be a training dataset composed of n i.i.d. samples drawn from $\mathcal{D}$. As usual, we denote $L_{\mathcal{D}}(h) = E_{(x,y)\sim\mathcal{D}}[l(h(x), y)]$ and $L_S(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$ the true and empirical risks of a hypothesis h, respectively. For simplicity, let us denote by h_S the output of a learning algorithm when trained with dataset S.

An important property of learning algorithms is their ability to generalize, i.e., the true and empirical risks of the output hypothesis should be close in expectation. Formally, we say that a learning algorithm $\mathcal{A}$ ϵ-generalizes in expectation if

$$|E_S[L_S(h_S) - L_{\mathcal{D}}(h_S)]| < \epsilon \ . \tag{3}$$

An interesting connection arises when we investigate the *stability* of a learning algorithm. Formally, we call a learning algorithm *ϵ-uniformly stable* if $\forall S, S'$ datasets of size n that differ in at most one example we have

$$\sup_{(x,y)} l(h_S(x), y) - l(h_{S'}(x), y) < \epsilon \ . \tag{4}$$

Notations: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n), (\widetilde{x}_1, \widetilde{y}_1), \dots, (\widetilde{x}_n, \widetilde{y}_n)$ are $2n$ independently sampled training examples. We define $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$, $\widetilde{S} = \{(\widetilde{x}_1, \widetilde{y}_1), \dots, (\widetilde{x}_n, \widetilde{y}_n)\}$ and $S^{(i)} = \{(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (\widetilde{x}_i, \widetilde{y}_i), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n)\}$.
Prove that:

1. (5 pts) $L_{\mathcal{D}}(h_S) = E_{\widetilde{S}}[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\widetilde{x}_i), \widetilde{y}_i)]$.
2. (8 pts) $E_{S,\widetilde{S}}[l(h_S(\widetilde{x}_i), \widetilde{y}_i)] = E_{S,S^{(i)}}[l(h_{S^{(i)}}(x_i), y_i)]$.
3. (12 pts) An ϵ-uniformly stable learning algorithm ϵ-generalizes in expectation.

Solution:

1. Note that since $\tilde{S}$ is composed of n i.i.d. samples $L_{\mathcal{D}}(h_S) = E_{(\tilde{x_i},\tilde{y_i})\sim\mathcal{D}}[l(h_S(\tilde{x_i}), \tilde{y_i})]$ for all i. Thus, by linearity of expectation $L_{\mathcal{D}}(h_S) = E_{\tilde{S}}[\frac{1}{n}\sum_{i=1}^{n} l(h_S(\tilde{x_i}), \tilde{y_i})]$.

2. $E_{S,\tilde{S}}[l(h_S(\tilde{x_i}), \tilde{y_i})] = E_{S,(\tilde{x_i},\tilde{y_i})}[l(h_S(\tilde{x_i}), \tilde{y_i})] =$

   (since $(x_1, y_1), \ldots, (x_n, y_n), (\tilde{x_i}, \tilde{y_i})$ *are i.i.d. we can interchange* (x_i, y_i) *with* $(\tilde{x_i}, \tilde{y_i})$ *)*

   $= E_{S^{(i)},(x_i,y_i)}[l(h_{S^{(i)}}(x_i), y_i)]$

3. $$\begin{aligned}
&|E_S[L_S(h_S) - L_{\mathcal{D}}(h_S)]| \overset{(1)}{=} |E_S\left[L_S(h_S) - E_{\tilde{S}}\left[\tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} l(h_S(\tilde{x_i}), \tilde{y_i})\right]\right]| = \\
&= |E_S\left[L_S(h_S)\right] - E_{S,\tilde{S}}\left[\tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} l(h_S(\tilde{x_i}), \tilde{y_i})\right]| = \\
&= |E_S\left[L_S(h_S)\right] - \tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} E_{S,\tilde{S}}\left[l(h_S(\tilde{x_i}), \tilde{y_i})\right]| \overset{(2)}{=} \\
&= |E_S\left[L_S(h_S)\right] - \tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} E_{S^{(i)},(x_i,y_i)}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| = \\
&= |E_S\left[\tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} l(h_S(x_i), y_i))\right] - \tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} E_{S,S^{(i)}}\left[l(h_{S^{(i)}}(x_i), y_i)\right]| = \\
&= |\tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} E_{S,S^{(i)}}\left[l(h_S(x_i), y_i)) - l(h_{S^{(i)}}(x_i), y_i)\right]| \overset{(\text{ }\epsilon\text{-uniform stability})}{\leq} \\
&\leq \tfrac{1}{n}{\textstyle\sum_{i=1}^{n}} \epsilon = \epsilon
\end{aligned}$$

Problem 4. *VC dimension of union* (20 pts) Let $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_r$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = \max_i \text{VCdim}(\mathcal{H}_i)$ and assume that $d > 2$.
Prove that:

1. (13 pts) $\text{VCdim}(\bigcup_{i=1}^r \mathcal{H}_i) \leq \frac{4d}{\log(2)} \log\left({}^{2d}/_{\log(2)}\right) + \frac{2\log(r)}{\log(2)}$.
   Hint: Use Sauer's lemma for bounding the growth function and the inequality

   "Let $a \geq 1$ and $b > 0$. If $x \leq a\log(x) + b$ then $x \leq 4a\log(2a) + 2b$."

2. (7 pts) For $r = 2$ the bound can be strengthen to $\text{VCdim}(\mathcal{H}_1 \cup \mathcal{H}_2) \leq 2d + 1$.
   Hint: $\sum_{i=0}^k \binom{k}{i} = 2^k$

Solution:

1. Let $\mathcal{H} = \bigcup_{i=1}^r \mathcal{H}_i$. By definition of the growth function we have $\tau_{\mathcal{H}}(m) \leq \sum_{i=1}^r \tau_{\mathcal{H}_i}(m)$ for any set of m points. If $k > d+1$ points are shattered by $\mathcal{H}$ then $2^k = \tau_{\mathcal{H}}(k) \leq \sum_{i=1}^r \tau_{\mathcal{H}_i}(k) \leq rk^d$, where the last inequality follows directly from Sauer's lemma. Taking the logarithm on both sides and using the inequality yields

   $$k \leq \frac{4d}{\log(2)} \log\left(\frac{2d}{\log(2)}\right) + 2\frac{\log(r)}{\log(2)} .$$

   Note that this inequality is trivially satisfied if $k \leq d+1$.

2. Assume that $k \geq 2d+2$. It is enough to prove that $\tau_{\mathcal{H}_1 \cup \mathcal{H}_2}(k) < 2^k$.

$$\begin{aligned}
\tau_{\mathcal{H}_1 \cup \mathcal{H}_2}(k) \leq \tau_{\mathcal{H}_1}(k) + \tau_{\mathcal{H}_2}(k) \leq \sum_{i=0}^d \binom{k}{i} + \sum_{i=0}^d \binom{k}{i} = \\
= \sum_{i=0}^d \binom{k}{i} + \sum_{i=0}^d \binom{k}{k-i} = \sum_{i=0}^d \binom{k}{i} + \sum_{i=k-d}^k \binom{k}{i} \leq \\
\leq \sum_{i=0}^d \binom{k}{i} + \sum_{i=d+2}^k \binom{k}{i} < \sum_{i=0}^d \binom{k}{i} + \sum_{i=d+1}^k \binom{k}{i} = \\
= \sum_{i=0}^k \binom{k}{i} = 2^k
\end{aligned}$$

Lemma. ***(Sauer-Shelah-Perles)*** *Let $\mathcal{H}$ be a hypothesis class with $VCdim(H) \leq d < \infty$ and growth function $\tau_{\mathcal{H}}$. Then, for all m, $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^d \binom{m}{i}$. In particular, if $m > d+1$ and $d > 2$ then $\tau_{\mathcal{H}}(m) < m^d$.*

Problem 5. *Short problems* (10 pts)

(i) (5 pts) You have lots and lots and lots of data. You use a neural net with a single hidden layer and run stochastic gradient descent. What theoretical framework(s) will likely give you meaningful insights for this situation? Explain why.

(a) NTK as discussed in the course

(b) mean field model as discussed in the course

(c) basic learning theory generalization bounds

(ii) (5 pts) Assume that you are in a scenario where the mean field model that we discussed in the course applies. Can you use the machinery discussed in the paper by Montanari to optimize and predict the performance of an actual system? What, if any, are the remaining problems. Write down a few (and we mean a few) sentences to discuss.

Solution:

(i) (5 pts) Since you have so much data the mean field model is likely going to predict your performance correctly. The NTK applies even for a fixed amount of data but it only applies in the setting of vanishing learning rate (and the width of the network should be large). Since you have so much data the basic generalization bound will also tell you that if you chose according to the empirical mean you will likely choose close to an optimal hypothesis. But you still need to argue that the stochastic gradient will in fact do well in this scenario.

(ii) (5 pts) The main problem is that there is no easy and efficient way to compute the solution of the associated differential equation. In fact, solving such types of differential equations is typically done by running stochastic gradient descent! :-) So this framework can be used to discuss convergence and other theoretical questions but currently cannot be used to predict the performance or to optimize the parameters of the system.