Prof. J. M. Buhmann

Final Exam

August 28th, 2010

First and Last name: ____________________

ETH number: ____________________

Signature: ____________________

General Remarks

- Please check that you have all 24 pages of this exam.
- Remove all material from your desk which is not permitted by the examination regulations.
- Fill in your name and ETH number and sign the exam. Place your student ID on the desk.
- You have 120 minutes for the exam. There are five questions, where you can earn a total of 120 points. You don't need to score every point to earn the top grade.
- Write your answers directly on the exam sheets. If you need more space, put your name and ETH number on top of each supplementary sheet.
- Answer the questions in English. Do not use a pencil or red color pen.
- You may provide at most one valid answer per question. Invalid solutions must be canceled out clearly.

	Topic	Max. Points	Points Achieved	Visum
1	Clustering	30		
2	Dimensionality Reduction	30		
3	Role Mining	20		
4	Deterministic Annealing	10		
5	Sparse Coding	30		
Total		120		

Grade: ....

Question 1: Clustering (30 pts.)

a) We have N data points in D dimensions. Perform the first iteration of the K-means algorithm with the Euclidean distance by hand. Assume $K = 2$ clusters and that the first and second centroid are initialized to the first and second data point, respectively. Give the answer in the matrix factorization format: $\mathbf{X} \approx \mathbf{UZ}$, with the data points $\mathbf{X} \in \mathbb{R}^{D \times N}$, the centroids $\mathbf{U} \in \mathbb{R}^{D \times K}$ and the assignments $\mathbf{Z} \in \{0, 1\}^{K \times N}$. The data is given by

$$\mathbf{X} = \begin{bmatrix} 1 & 3 & 2 & 0 & 2 \\ 0 & 3 & 3 & 0 & 2 \end{bmatrix}.$$

Compute the matrices $\mathbf{U}^{(0)}, \mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(1)}$.
Answer:

i) Initialization centroids.

$$\mathbf{U}^{(0)} = \begin{bmatrix} \qquad\qquad \\ \\ \end{bmatrix}$$

ii) Assignment step.

$$\mathbf{Z}^{(1)} = \begin{bmatrix} \qquad\qquad\qquad\qquad\qquad\qquad \\ \\ \end{bmatrix}$$

iii) Update centroids.

$$\mathbf{U}^{(1)} = \begin{bmatrix} \qquad\qquad \\ \\ \end{bmatrix}$$

You may use fractions or round-off to the first decimal. **6 pts.**

b) Derive the centroids update for the k-means algorithm. Start your derivation by considering the cost $J(\mathbf{U}, \mathbf{Z})$ of a clustering

$$J(\mathbf{U}, \mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2.$$

Here $\mathbf{u}_k$ denotes the k-th centroid and $z_{k,n}$ the assignment of data point $\mathbf{x}_n$ to the k-th cluster. We expect you to write down all intermediate steps of the centroid update derivation.

Answer:

5 pts.

In the lecture we have discussed the Gaussian mixture model (GMM)

$$p_{GMM}(x|\boldsymbol{\pi}, \boldsymbol{\mu}, \sigma) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \sigma),$$

in which $\mathcal{N}(x|\mu, \sigma)$ is the 1-D Gaussian distribution with mean μ and standard deviation σ,

$$\mathcal{N}(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Let us now consider the problem of *sampling from the GMM*. Sampling describes the process of generating (or drawing) data points from a given distribution, which in our case is given by the GMM from above. Sampling a data point from the GMM is done in two steps.

c) Describe the two steps in words.

Answer:

2 pts.

Having identified the two parts, we are now interested in an algorithm for sampling from the Gaussian mixture model distribution.

d) Give a detailed Matlab implementation for sampling a data point from the 1-D Gaussian mixture model. Assume that the means of the different components are given by $[\mu_1, \dots, \mu_K]$ in the vector mus, and that all components have the same standard deviation sigma. Furthermore, you are given the mixture weights $[\pi_1, \dots, \pi_K]$ in the vector pis.

Hints:

- You can use a function x = rand_gaussian(mu, sigma) that generates a data point from a Gaussian distribution with mean mu and standard deviation sigma.
- You will need to use the function x = rand(), which generates a random number uniformly in $(0, 1]$.

Answer:

```
function x = sample_gmm(pis, mus, sigma)
```

8 pts.

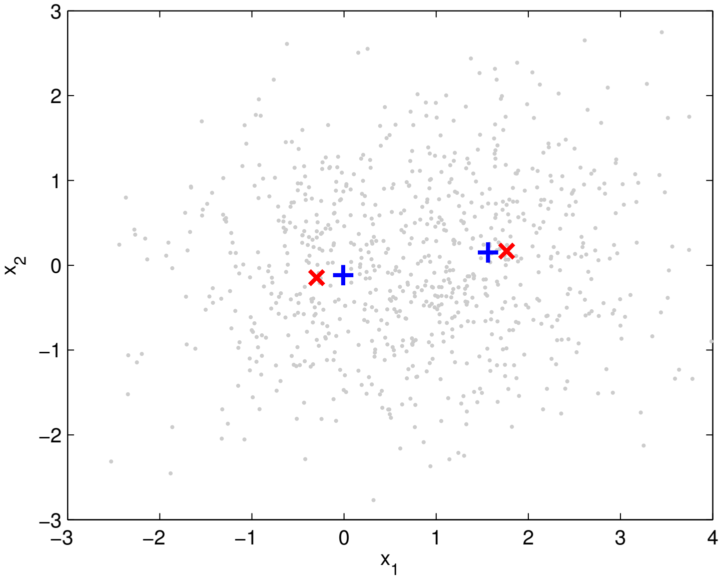
Imagine you are given a box of fair dice: 20% of the dice are four sided showing the numbers $1, \ldots, 4$, and 80% of the dice are six sided with numbers $1, \ldots, 6$. You now pick a die in this box at random and throw the die that you just chose and write down the number shown on the die. Finally you put the die back into the box.

e) Give a concise mathematical expression for $P(X = n)$, where X is the random variable that denotes the outcome of the process described above and n is an integer in $1, \ldots, 6$. Do not compute the values of all these events, but only give a symbolic expression.

   Answer:

6 pts.

f) Below you see the centroids of the solution of two clustering algorithms. One solution is computed by the k-means algorithm (hard assignments), one by the Gaussian mixture model (soft assignments). Which estimated centroids correspond to which clustering algorithm? We expect you to give a reasoning for your solution (no points if no explanation is given).



Answer:

Centroid-pair $+$ corresponds to:

Centroid-pair $\times$ corresponds to:

Explanation:

3 pts.

Question 2: Dimensionality Reduction (30 pts.)

Singular Value Decomposition

We are given a user-item matrix $\mathbf{A}$ of 6 users and 7 items, where $a_{i,j}$ contains a value between 0 (dislike) and 10 (like) indicating the preference of user i for item j. We apply SVD on $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ and obtain the following (rounded off) matrices $\mathbf{U}$ and $\mathbf{V}^\top$.

$$\mathbf{U} = \begin{bmatrix} 0.4 & 0.2 & 0.5 & 0.5 & -0.3 & -0.4 \\ 0.3 & -0.6 & -0.1 & -0.4 & -0.1 & -0.5 \\ 0.4 & 0.3 & -0.5 & -0.1 & -0.6 & 0.3 \\ 0.4 & 0.2 & 0.5 & -0.5 & 0.3 & 0.4 \\ 0.3 & -0.6 & -0.1 & 0.4 & 0.1 & 0.5 \\ 0.4 & 0.3 & -0.5 & 0.2 & 0.6 & -0.3 \end{bmatrix},$$

$$\mathbf{V}^\top = \begin{bmatrix} 0.40 & 0.40 & 0.37 & 0.35 & 0.40 & 0.35 & 0.37 \\ 0.39 & 0.43 & -0.38 & -0.40 & 0.43 & -0.33 & -0.27 \\ 0.01 & 0.05 & 0.47 & 0.46 & 0.05 & -0.53 & -0.53 \\ -0.56 & 0.29 & -0.17 & 0.15 & 0.29 & 0.49 & -0.47 \\ 0.12 & -0.06 & 0.62 & -0.63 & -0.06 & 0.31 & -0.32 \\ -0.60 & 0.26 & 0.30 & -0.28 & 0.26 & -0.39 & 0.43 \\ 0 & -0.71 & 0 & 0 & 0.71 & 0 & 0 \end{bmatrix}$$

1. Please write down the corresponding matrix $\mathbf{D}$. To help you, we reveal the values of the following elements: $d_{1,1} = 45$, $d_{2,2} = 20$, $d_{3,3} = 17$, $d_{4,4} = 1$, $d_{5,5} = 0.7$, and $d_{6,6} = 0.2$.
Answer:

2 pts.

2. Given the factorization $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, write down the equation for $\mathbf{A}$ as a full sum of outer products.
   Answer:

2 pts.

3. We would like to represent the data with fewer dimensions, resulting in an approximation. We compute a good approximation for $\mathbf{A}$ by selecting $k < 6$ dimensions to keep. Choose the dimensions you want to keep and write down the equation for your approximation $\tilde{\mathbf{A}}$ as a sum of outer products of these selected dimensions. Explain your choice.
   Answer:

4 pts.

4. For your choice above, box out the components in $\mathbf{U}$, $\mathbf{V}^\top$ and $\mathbf{D}$ that we would like to keep.
   Answer: *Box out in the previous page. Do not forget to box* $\mathbf{D}$ *as well!*

4 pts.

5. Please write down the numerical value of the approximation error (under the Euclidean matrix norm) of your solution.
   Answer:

1 pt

6. How do we deduce the preference of user i for item j in your approximation $\tilde{\mathbf{A}}$? Write down the equation.
   Answer:

2 pts.

7. Write down the numerical sums for computing the preference of user 5 for item 3. Roughly estimate its value and tell us if user 5 likes item 3 (i.e., is this value closer to 10 or 0)? (Do this easily by rounding-off all numbers in the numerical sum to one significant value.)
   Answer:

3 pts.

8. Whoops! One item was mistakenly included twice in the original matrix $\mathbf{A}$ before performing the SVD (i.e., we actually only have 6 items, and 1 item was replicated). Which two items are identical? Explain your choice.
   Answer:

4 pts.

Principal Component Analysis

We are given a data set $\mathbf{X} \in \mathbb{R}^{D \times N}$ of N samples of dimension D and its SVD decomposition $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Furthermore, we state that $\mathbf{X}$ is zero-mean, i.e., the mean of the samples along each dimension is 0.

We would like to perform a dimensionality reduction on $\mathbf{X}$ using PCA.

1. In PCA, we need to perform an eigenvalue decomposition on a particular quantity $\mathbf{H}$. What is this quantity $\mathbf{H}$, and write it in terms of $\mathbf{X}$.
   Answer:

   2 pts.

2. We can use the SVD solution to compute the PCA. Show this by first plugging in the SVD factorization $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ into your definition of $\mathbf{H}$. Use the fact that $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$ (where $\mathbf{I}$ is the identity matrix) to simply your equation.
   Answer:

   4 pts.

3. Second, explain how your answer to the previous question can be used to obtain the PCA for $\mathbf{X}$: what are the eigenvalues and respective eigenvectors of $\mathbf{H}$?
   Answer:

   2 pts.

Question 3: Role Mining (20 pts.)

Permission assignments via roles Assume a binary user-permission matrix $\mathbf{X}$ with D permissions and N users, which we factorize into $\mathbf{U}$ (the matrix of roles) and $\mathbf{Z}$ (the assignment matrix of roles to users). There are various evaluation criteria for role mining solutions.

a) State the formula and a short description for the **deviation** (mean Hamming distance). Your explanation should include what a low (and a high) deviation implies.
Answer:

Formula:

Explanation:

4 pts.

b) The **generalization error** of a set of roles to a new user-permission matrix $\mathbf{X}'$ is computed as follows: For each new user i, the optimal role assignment $\hat{\mathbf{z}}_{\cdot,i}$ is determined based on a subset $\mathcal{D}^* \subset \{1, \ldots, D\}$ of the permissions:

$$\hat{\mathbf{z}}_{\cdot,i} := \arg\min_{\mathbf{z}_{\cdot,i}} \left\{ \sum_{d \in \mathcal{D}^*} \left| x'_{d,i} - \mathbf{u}_{d,\cdot} \otimes \mathbf{z}_{\cdot,i} \right| \right\}$$

The generalization error is then

$$G := \frac{1}{N} \sum_{i=1}^{N} \frac{1}{D} \left|\left| \mathbf{x}'_{\cdot,i} - \mathbf{U} \otimes \hat{\mathbf{z}}_{\cdot,i} \right|\right|_1$$

1. What does a high **generalization error** imply?
Answer:

1 pt.

2. What does the **generalization error** and the **deviation** have in common? How do they differ?
   Answer:

1 pt.

3. Let a noisy dataset $\tilde{\mathbf{X}}$ be given. Is the existence of a set of roles with generalization error $G = 0$ possible? Explain your answer.
   Answer:

1 pt.

c) Let the three roles $r_1 = \{p_1, p_3, p_4\}$, $r_2 = \{p_2, p_3\}$, and $r_3 = \{p_1, p_3\}$ be given, where p_d is the d^{th} permission.

1. A user is assigned the role combination $\{r_1, r_2\}$. What are his permissions? Answer:

3 pts.

2. How many combinations of the roles r_1, r_2, and r_3 contain p_1? How many combinations do not contain p_1? (Assigning no role does not count as a combination.)
   Answer:

5 pts.

3. The assignment vector $\mathbf{z}_{\cdot,2}$ encodes that user number two is assigned the roles r_4, r_5, and r_6. Let $t_{d,k}$ be the probability that role k contains permission d. Derive the probability $p(x_{5,2} = 1|\mathbf{z}_{\cdot,2})$ that the user has permission 5 given his roles. Compute the magnitude of this probability for $t_{5,4} = 0.9$, $t_{5,5} = 0.7$, and $t_{5,6} = 0.5$. You can assume that $t_{5,4}$, $t_{5,5}$, and $t_{5,6}$ are independent.
   Answer:
   $p(x_{5,2} = 1|\mathbf{z}_{\cdot,2}) =$

5 pts.

Question 4: Deterministic Annealing (10 pts.)

Let the data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ consist of N data items. There are K clusters numbered from 1 to K. Each cluster k is described by its parameters θ_k. A given clustering model specifies *costs* $R_{n,k}$ for assigning data item n to cluster k, for $n = 1, \ldots, N$ and $k = 1, \ldots, K$.

a) In the following, you are given pseudo-code for deterministic annealing. However, the lines are mixed up. Please sort the lines such that the pseudo-code is correct. You don't have to write down the corrected code, just give the line numbers in correct order.

1. M-step: For $k = 1, \ldots, K$, optimize the parameters θ_k given the current values $\gamma_{n,k}$
2. **end while**
3. E-step: Given the current parameter values θ_k, compute $R_{n,k}$ and $\gamma_{n,k} = \frac{\exp\left[-\frac{1}{T}R_{n,k}\right]}{\sum_k \exp\left[-\frac{1}{T}R_{n,k}\right]}$ for all n and k.
4. Decrease $T : T \leftarrow \alpha \cdot T$, $\alpha < 1$
5. **while** not converged
6. Determine whether the algorithm has converged
7. Initialize: $T = T_{start}$, converged = false, θ_k = random initialization

Answer:

6 pts.

b) What is the effect of the parameter T in deterministic annealing? Discuss in particular the limits $T \to \infty$ and $T \to 0$.

Answer:

4 pts.

Question 5: Sparse Coding (30 pts.)

Orthogonal Matrices and Bases

a) What are the *three* properties that $\mathbf{U}$ has to satisfy, such that it is an *orthogonal* matrix?
Answer:

3 pts.

b) Here is a proof for a very useful property of orthogonal matrices. Given an orthogonal matrix $\mathbf{U}$ and a vector $\mathbf{x}$ (both of proper size), it holds that

$$\begin{aligned}
\left\|\mathbf{U}^\top\mathbf{x}\right\|_2^2 &= \left(\mathbf{U}^\top\mathbf{x}\right)^\top\left(\mathbf{U}^\top\mathbf{x}\right) && \left(\qquad\qquad\qquad\qquad\qquad\qquad\right)\\
&= \mathbf{x}^\top\mathbf{U}\mathbf{U}^\top\mathbf{x} && \left(\qquad\qquad\qquad\qquad\qquad\qquad\right)\\
&= \mathbf{x}^\top\mathbf{I}\mathbf{x} && \left(\qquad\qquad\qquad\qquad\qquad\qquad\right)\\
&= \left\|\mathbf{x}\right\|_2^2, && \left(\qquad\qquad\qquad\qquad\qquad\qquad\right)
\end{aligned}$$

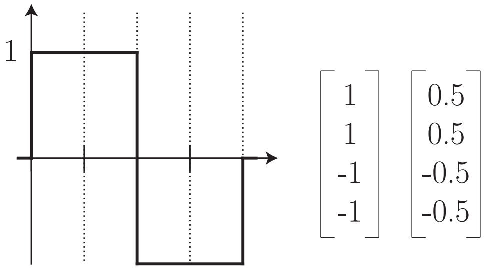
where $\mathbf{I}$ is the *identity matrix*. First, for every line of the proof, write down in the space next to it why this equality holds. Second, what is the property that we have proven here? Give your answer in the form of a theorem.
Answer:

5 pts.

c) The *Haar wavelets* $\mathbf{H} \in \mathbb{R}^{4\times 4}$ form an orthogonal basis of $\mathbb{R}^4$. In the upper left figure, the mother wavelet $\mathbf{h}_1 \in \mathbb{R}^4$ is shown both as a graph and as a column vector. Complete the Haar basis using *dilation* and *translation* operations, i.e. add the remaining three basis column vectors.

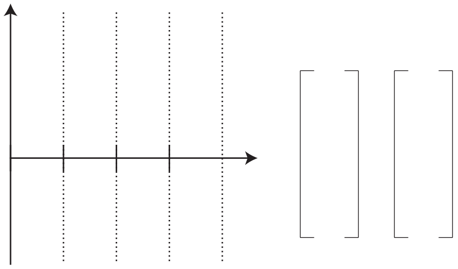
Finally, normalize each vector to unit Euclidean length, such that we have an *orthonormal* basis. Use the approximation $1/\sqrt{2} \approx 0.7$ in your calculations.

Answer:

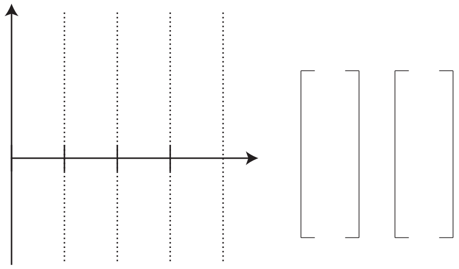


$$\begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix} \quad \begin{bmatrix} 0.5 \\ 0.5 \\ -0.5 \\ -0.5 \end{bmatrix}$$

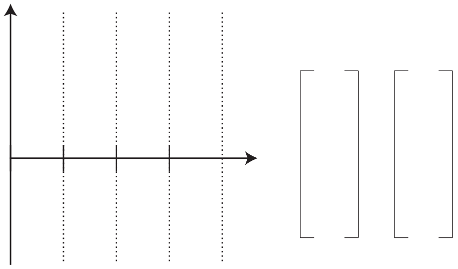
$\mathbf{h}_1$



$\mathbf{h}_2$



$\mathbf{h}_3$



$\mathbf{h}_4$

6 pts.

Compression by Sparse Coding

d) What are the *coding* and *thresholding* steps, to compress a signal $\mathbf{x} \in \mathbb{R}^4$ using the Haar basis $\mathbf{H}$?

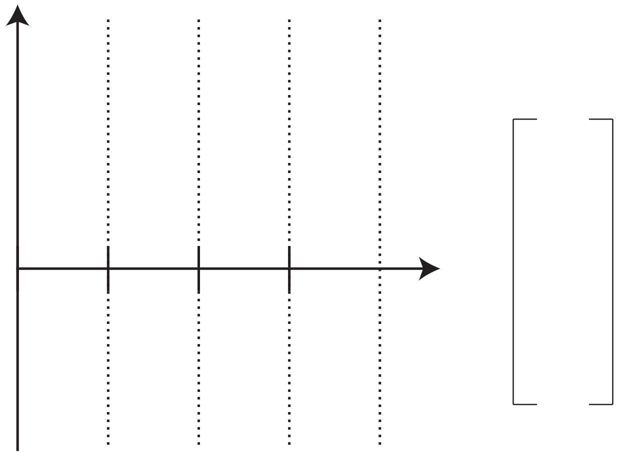
Answer:

2 pts.

e) What is the least compressible signal $\mathbf{y} \in \mathbb{R}^4$, i.e. the signal that achieves the worst compression ratio vs. reconstruction error, using sparse coding in the Haar basis? Consider only signals which satisfy $\|\mathbf{y}\|_2 = 2$.

Draw the graph of $\mathbf{y}$, give its representation as a vector and argue why it is the least compressible signal.

Answer:



$\mathbf{y}$

5 pts.

Matching Pursuit

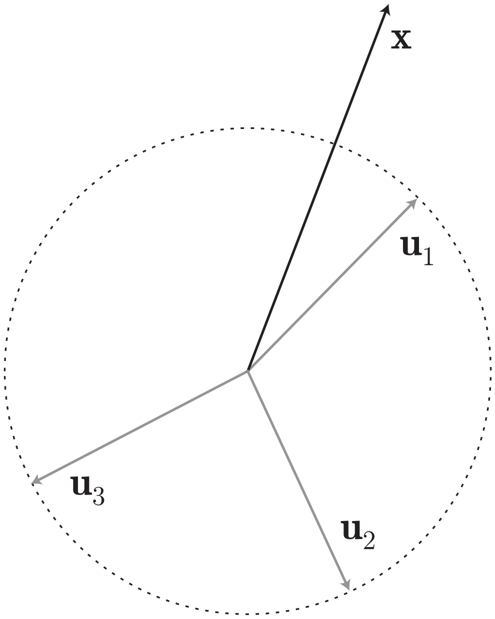
f) Here is an incomplete Matlab implementation of sparse coding by matching pursuit (MP). $\mathbf{U}$ is the dictionary, $\mathbf{x}$ the signal and k the desired cardinality of the coding $\mathbf{z}$:

```
function z = mp(U, x, k)

z = zeros(size(U,2), 1);
res = x;
card = 0;
while card < k

    [~, cur] = max( _____________ );
    coef = U(:,cur)'*res;

    res = res - _______________ ;
    z(cur) = z(cur) + coef;
    card = sum( z ~= 0 );
end
end
```

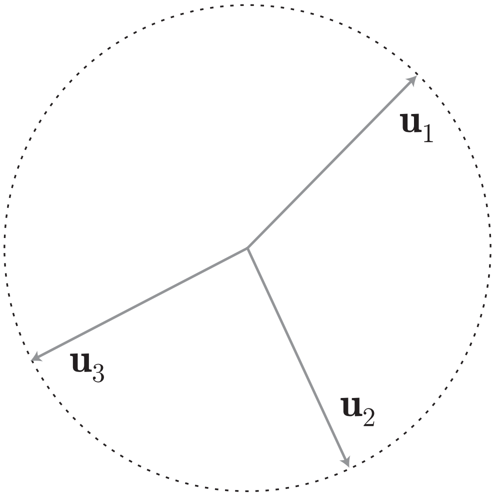
Complete the two blanks.

4 pts.

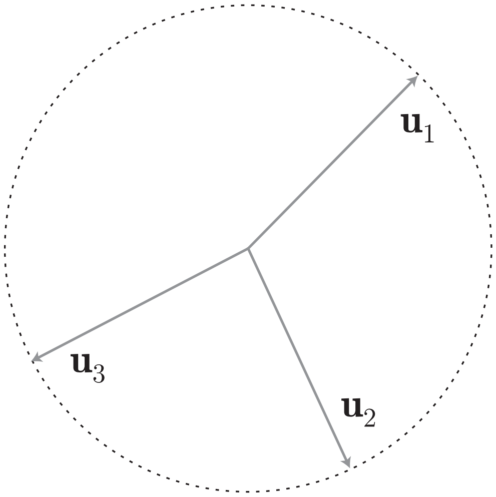
g) Visualize the MP algorithm on the following example, using $k = 2$. For each iteration of the while loop, illustrate the **atom choice** (circle the correct $\mathbf{u}_k$), **projection** (solid line) and **residual** (dashed line) computation steps. In the final figure, draw in the approximate reconstructed signal $\tilde{\mathbf{x}} = \mathbf{U}\mathbf{z}$.
Note: You may draft your solution in the draft sheet.
Answer:



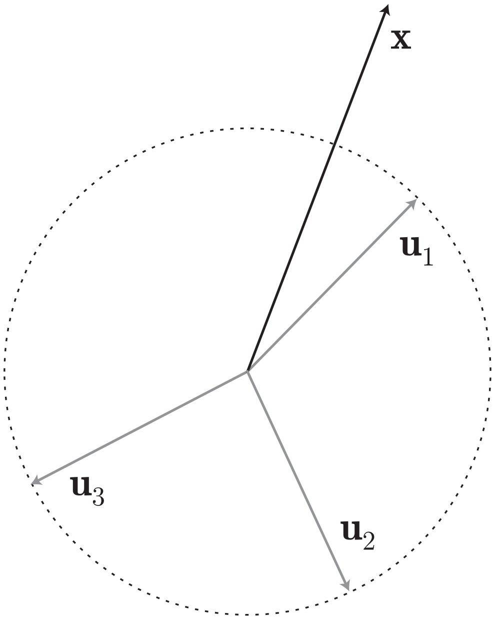
(1) First Iteration



(2) Second Iteration



(3) Third Iteration



(4) Approximate Reconstructed Signal

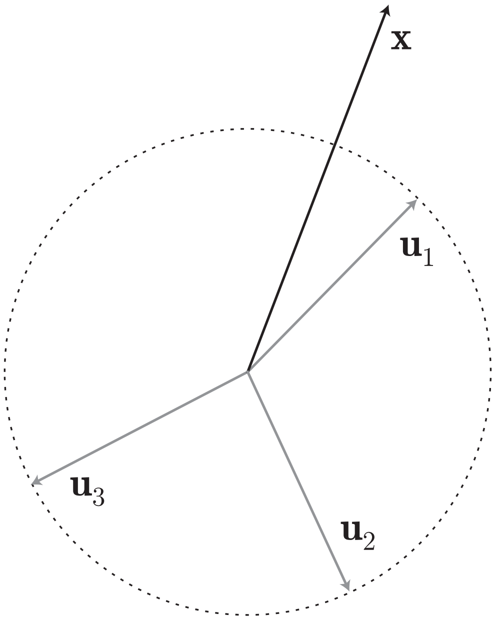
5 pts.

Supplementary Sheet

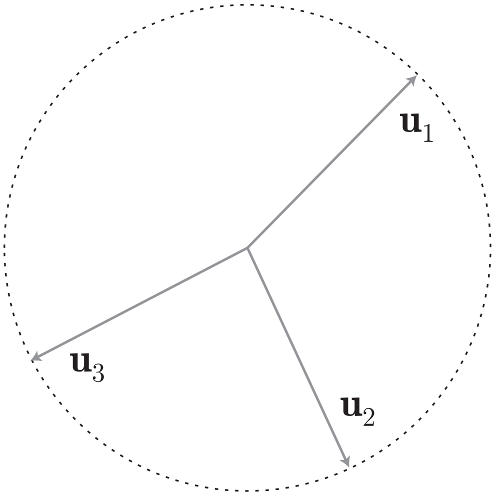
Supplementary Sheet

Supplementary Sheet

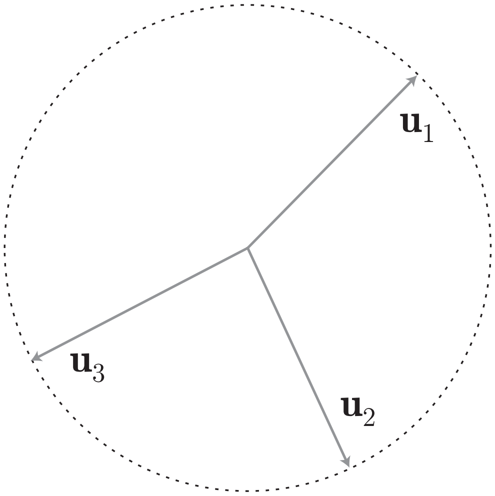
Draft Sheet for Question 5 Part g: This page will not be corrected.



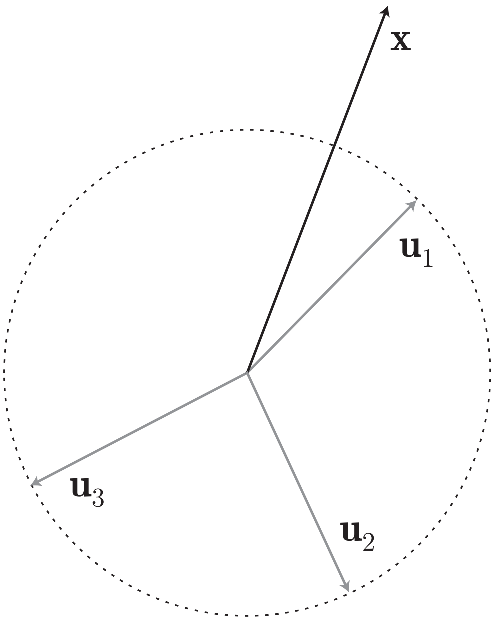
(1) First Iteration



(2) Second Iteration



(3) Third Iteration



(4) Approximate Reconstructed Signal

Notation

- $\mathbf{x}$ is a column vector
- $\top$ denotes the *transpose* operator, so $\mathbf{x}^\top$ is a row vector
- The elements of a vector are denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_D)^\top$
- $\mathbf{U}$ is a matrix:
  - $\mathbf{u}_k$ or $\mathbf{u}_{\cdot,k}$ is the k-th column of $\mathbf{U}$
  - $\mathbf{u}_d^\top$ or $\mathbf{u}_{d,\cdot}$ is the d-th row of $\mathbf{U}$
  - $u_{d,k}$ is the element in the d-th row and k-th column of $\mathbf{U}$

Please use this notation when answering questions - except the bold font style, of course.