ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computational Intelligence Lab

2017

Prof. T. Hofmann

Final Exam

21st August 2017

First and Last name: _____

Student ID (Legi) Nr: _____

Signature: _____

General Remarks

- Please check that you have all 22 pages of this exam.
- There are 120 points, and the exam is 120 minutes. **Don't spend too much time on a single question!** The maximum of points is not required for the best grade!
- Remove all material from your desk which is not permitted by the examination regulations.
- Write your answers directly on the exam sheets. If you need more space, make sure you put your **student-ID**-number on top of each supplementary sheet.
- Immediately inform an assistant in case you are not able to take the exam under regular conditions. Later complaints are not accepted.
- Attempts to cheat/defraud lead to immediate exclusion from the exam and can have judicial consequences.
- Please use a black or blue pen to answer the questions.
- Provide only one solution to each exercise. Cancel invalid solutions clearly.

	Topic	Max. Points	Points Achieved	Visum
1	Dimensionality Reduction	30		
2	Optimization, NMF, pLSA and Word Embedding	30		
3	Neural networks and clustering	30		
4	Sparse Coding and Dictionary Learning	30		
Total		120		

Grade: .....

1 Dimensionality reduction (30 pts)

1.1 PCA on an Ellipse

Our goal is to apply PCA to a set of points sampled from an ellipse. Let $a > b$ be two positive real numbers. Assume we have the following $4n$ 2-dimensional data points: $\mathcal{P} := \{P_k\}_{k=0,\dots,4n-1}$, $P_k = \left(a \cdot \cos\frac{2k\pi}{4n}, b \cdot \sin\frac{2k\pi}{4n}\right)$.

Reminder about trigonometric relations:

- $\sin(x) = -\sin(2\pi - x) = \sin(\pi - x) = -\sin(\pi + x)$
- $\cos(x) = \cos(2\pi - x) = -\cos(\pi - x) = -\cos(\pi + x)$
- $\cos(2x) = 2\cos^2(x) - 1 = 1 - 2\sin^2(x)$

a) Show that the given set of points $\mathcal{P}$ satisfies the ellipse equation $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$.

1 pts ☐

. . . . . .

. . . . . .

b) Compute the mean vector of the data $\mathcal{P}$.

2 pts ☐

. . . . . .

. . . . . .

c) Compute the covariance matrix of the data $\mathcal{P}$.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

d) Compute the first principal component of $\mathcal{P}$ as well as the corresponding eigenvalue.

1 pts ☐

. . . . . .

. . . . . .

e) Compute the new coordinates of $\mathcal{P}$ after applying PCA to reduce the dimension to $d = 1$.

1 pts ☐

. . . . . .

. . . . . .

f) Compute the reconstruction error of PCA after projection to a 1-dimensional space.

1 pts ☐

. . . . . .

. . . . . .

1.2 PCA in 3D

We are given a set $\mathcal{P}'$ of $2n+1$ 3-dimensional data points with coordinates $(2i, -5i, 3i)$, where $i \in \{-n, \ldots, 0, \ldots, n\}$.

a) Compute the covariance matrix of $\mathcal{P}'$. Hint: $\sum_{i=1}^{n} i^2 = \frac{1}{6}n(n+1)(2n+1)$

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

b) Compute the first principal component and its corresponding eigenvalue.

2 pts ☐

. . . . . .

. . . . . .

c) Compute the coordinates of the set of points after applying PCA to reduce the data dimension from $d = 3$ to $d = 1$.

2 pts ☐

. . . . . .

. . . . . .

d) Compute the reconstruction error of PCA in this case.

1 pts ☐

. . . . . .

. . . . . .

1.3 SVD for Solving Linear Systems

We are given a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$. We would like to find the minimum-norm least squares solution to a linear system $\mathbf{A}x = b$, that is, to find the vector x that achieves:

$$\min_x \|\mathbf{A}x - b\|_2.$$

Assume we can compute the SVD of matrix $\mathbf{A}$ as $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, where $\Sigma = diag(\sigma_1, \ldots \sigma_r, 0, \ldots 0)$. We also introduce the matrix $\Sigma^\dagger := diag(\frac{1}{\sigma_1}, \ldots \frac{1}{\sigma_r}, 0, \ldots 0)$.

a) Prove that the optimal shortest (minimum 2-norm) vector x^* minimizes $\|\Sigma\mathbf{V}^\top x - \mathbf{U}^\top b\|_2$

2 pts ☐

. . . . . .

. . . . . .

b) Prove that the optimal shortest (minimum 2-norm) vector y^* that achieves for a fixed vector c:

$$\min_y \|\Sigma y - c\|_2$$

is unique and is given by the formula $y^* = \Sigma^\dagger c$

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

c) Prove that the optimal vector x^* for the original problem (question a)) is unique and is given by the formula $x^* = \mathbf{V}\Sigma^\dagger\mathbf{U}^\top b$.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

1.4 SVD Properties

Let the SVD of a squared matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ be $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, where the matrix $\mathbf{U}$ contains the left singular vectors u_i as columns, $\mathbf{V}$ contains the right singular vectors v_i as columns and Σ contains the singular values σ_i on its diagonal in non-increasing order, i.e. $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$.

a) Prove that $\mathbf{A} = \sum_{i=1}^{n} \sigma_i u_i v_i^\top$.

2 pts ☐

. . . . . .

. . . . . .

b) Prove that $\mathbf{A}\mathbf{A}^\top$ and $\mathbf{A}^\top\mathbf{A}$ have exactly the same set of non-zero eigenvalues.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

c) Let k be a positive integer and $\mathbf{A}$ be a square symmetric matrix. Compute $\mathbf{A}^k$ based on the SVD decomposition of $\mathbf{A}$.

2 pts ☐

. . . . . .

. . . . . .

2 Optimization, NMF, pLSA and Word Embedding

(30 pts)

2.1 Optimization

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ we say that $\mathbf{x}^\star$ is a global minimum of f if:

$$\forall\ \mathbf{x} \in \mathbb{R}^n \quad f(\mathbf{x}^\star) \leq f(\mathbf{x})$$

On the other hand we say that $\bar{\mathbf{x}}$ is a local minimum of f if:

$$\exists \epsilon > 0 \text{ such that } \forall\ \mathbf{x} \in \mathcal{B}_\epsilon(\bar{\mathbf{x}}) \quad f(\bar{\mathbf{x}}) \leq f(\mathbf{x})$$

where $\mathcal{B}_\epsilon(\bar{\mathbf{x}})$ is a ball of radius ϵ centered at $\bar{\mathbf{x}}$ (i.e. $\mathcal{B}_\epsilon(\bar{\mathbf{x}}) = \{\mathbf{x} \mid \|\mathbf{x} - \hat{\mathbf{x}}\|_2 < \epsilon\}$)

a) Show that if f is convex then a local minimum is also a global minimum. (Hint: proof by contradiction)

3 pts

. . . . . .

. . . . . .

. . . . . .

. . . . . .

b) Let us consider the gradient descent update on the function f which we assume is convex and β-smooth with $\beta > 0$. We say that f is β-smooth if the following condition holds:

$$-\nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^\star) \leq -\frac{\|\nabla f(\mathbf{x})\|_2^2}{2\beta}$$

Let γ be a constant stepsize used for gradient descent. Show that, at any iteration $t \geq 0$ of gradient descent, if $\gamma \leq \frac{1}{\beta}$, then it holds that:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_2^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|_2^2$$

3 pts

. . . . . .

. . . . . .

. . . . . .

. . . . . .

c) Under the same assumption as the previous question, what conclusion can we draw if $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_2^2 = \|\mathbf{x}_t - \mathbf{x}^\star\|_2^2$?

2 pts ☐

. . . . . .

. . . . . .

d) We now assume that f is β-smooth but *not* convex and that it does not have any saddle point. What conclusion can we draw if $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|_2^2 = \|\mathbf{x}_t - \mathbf{x}^\star\|_2^2$ for $t > 0$?

2 pts ☐

. . . . . .

. . . . . .

e) Suppose we want to minimize a function $f(\mathbf{x})$ for which computing the full gradient is too computationally expensive. Therefore, we decide to compute the gradient only with respect to one coordinate $\mathbf{x}_k$ at time t. We therefore implement the following coordinate descent update, $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla_k f(\mathbf{x}_t)$, where $\nabla_k f(\mathbf{x}_t)$ is the gradient of f computed at $\mathbf{x}_t$ only with respect to a given coordinate $k \in [1, n]$. The function we are minimizing is:

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1, \qquad \mathbf{y} \in \mathbb{R}^p, \mathbf{A} \in \mathbb{R}^{p \times n}, \lambda > 0$$

where $\|\cdot\|_1$ is the L1 norm. Derive an expression for $\nabla_k f(\mathbf{x}_t)$. You can assume $\mathbf{x} \neq \mathbf{0}$.

2 pts ☐

. . . . . .

. . . . . .

2.2 pLSA and Word Embedding

Assume that we have a corpus containing the following 3 documents: $d_i, i = 1, 2, 3$.

- d_1 : The king built the castle
- d_2 : The king rode to the castle
- d_3 : The king likes the castle

There are in total $M = 7$ words w_j in the vocabulary: $\mathcal{V} = \{$the, king, built, castle, rode, to, likes$\}$. (capitalization insensitive)

a) The first step to construct a **pLSA** model is to summarize the corpus into a matrix of co-occurrence counts $\mathbf{X} = (x_{ij})$.

1) What is the meaning of an entry x_{ij}?

2) Complete the co-occurrence matrix $\mathbf{X}$ below.

3 pts

. . . . . .

	the	king	built	castle	rode	to	likes
d_1		1					0
d_2					1		
d_3		1			0		1

Table 1: Co-occurrence matrix $\mathbf{X}$

b) Suppose there are K topics, each identified with an integer $z \in \{1, ..., K\}$. In the context model, pLSA assumes that the occurrence of a word w in document d is modelled as

$$p(w|d) = \sum_{z=1}^{K} p(w|z)p(z|d).$$

1) What is the conditional independence assumption behind pLSA?

1 pts

. . . . . .

2) Let $u_{zi} := p(z|d_i), v_{zj} := p(w_j|z)$. $\mathbf{X} = \{x_{ij}\}$ is the co-occurrence matrix. Write down the log-likelihood $\mathcal{L}(\mathbf{U}, \mathbf{V})$ of the pLSA model, and the constraints on $\mathbf{U}, \mathbf{V}$.

1 pts

. . . . . .

. . . . . .

. . . . . .

3) The EM (expectation maximization) algorithm is used to maximize the log-likelihood of pLSA. Let q_{zij} be the variational parameters. Write down the lower bound of $\mathcal{L}(\mathbf{U}, \mathbf{V})$ and the update rule for q_{zij}.

2 pts

. . . . . .

c) We now consider the word embedding model known as **GloVe**. Let the window size be 1. Suppose both the word vocabulary and the context vocabulary are $\mathcal{V}$. Let the co-occurrence matrix be $\mathbf{N} = (n_{ij})$.

Complete the co-occurrence matrix $\mathbf{N}$ for the same corpus. (you just need to fill in entries in the upper triangle part of the matrix since $\mathbf{N}$ is symmetric here)

2 pts

	the	king	built	castle	rode	to	likes
the			1		0	1	1
king	x			0	1	0	1
built	x	x				0	0
castle	x	x	x		0		0
rode	x	x	x	x			0
to	x	x	x	x	x		0
likes	x	x	x	x	x	x	

Table 2: The co-occurrence matrix $\mathbf{N}$ of GloVe

d) Recall that the GloVe objective function is:

$$\mathcal{H}(\theta; \mathbf{N}) = \sum_{i,j} f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i, \mathbf{y}_j \rangle)^2,$$

where the bias terms are ignored for simplicity. $\mathbf{x}_i$ is the word embedding of the word i and $\mathbf{y}_j$ is the context embedding of the word j. $f : \mathbb{R} \to \mathbb{R}$ is a function that gives a weight to each pair (i, j) based on its number n_{ij}.

1) What does the $f(x)$ function in the GloVe model look like? (you can either draw it or write it down explicitly)

2) Briefly explain the purpose of this function.

3 pts

. . . . . .

. . . . . .

2.3 True/False Questions

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

6 pts

a) The GloVe objective with $f(x) = 1$ is equivalent to solving a matrix factorization problem

[ ] True [ ] False

b) Consider a non-negative matrix $\mathbf{X}$. We want to apply NMF to this matrix which consists in optimizing the following cost function,

$$\min_{\mathbf{U},\mathbf{V}} \quad J(\mathbf{U},\mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^\top\mathbf{V}\|_F^2.$$
$$\text{s.t. } u_{zi}, v_{zj} \geq 0, \forall(i,j,z)$$

This objective function is convex in $(\mathbf{U}^\top\mathbf{V})$.

[ ] True [ ] False

c) Consider using the Projected Alternating Least Squares algorithm to solve the NMF problem in Question b). The projection operation is used to ensure the non-negativity of $\mathbf{U}$ and $\mathbf{V}$.

[ ] True [ ] False

d) The EM algorithm used for maximizing the log-likelihood in pLSA is guaranteed to converge to the global optimum.

[ ] True [ ] False

e) pLSA and LDA have the same log likelihood

[ ] True [ ] False

f) The generative model of pLSA is the following:
For each document d in the corpus $\mathcal{D}$:

(a) sample a topic z from a prior $p(z)$

(b) sample a document d according to $p(d|z)$

(c) every word $w \in d$ is sampled according to $p(w|d,z) = p(w|z)$

[ ] True [ ] False

3 Neural networks and clustering (30 pts)

3.1 Neural Networks

a) We are given a neural network with one hidden layer with p neurons, followed by a sigmoid non-linearity σ, then a fully connected layer and finally a sigmoid that outputs a number between 0 and 1. The input is written $\mathbf{x} \in \mathbb{R}^n$, the hidden layer has weight matrix $\mathbf{W} \in \mathbb{R}^{p \times n}$ and bias $b \in \mathbb{R}^p$, and the fully connected layer is given by a vector $v \in \mathbb{R}^p$, so that the network computes $F(\mathbf{x}, \mathbf{W}, b, v) = \sigma(\sum_{i=1}^{p} v_i \sigma(\sum_{j=1}^{n} \mathbf{W}_{i,j} x_j + b_i))$. Compute the gradient of F w.r.t. $\mathbf{W}$. *Hint:* $\sigma'(x) = \sigma(x)(1 - \sigma(x))$.

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

b) Find a constant $L > 0$, as small as possible, such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$|F(\mathbf{x}, \mathbf{W}, b, v) - F(\mathbf{y}, \mathbf{W}, b, v)| \leq L \|\mathbf{x} - \mathbf{y}\|_2.$$

Express L in terms of the parameters of the network.
Hint: $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2, \quad \forall \mathbf{u}, \mathbf{v} \in \mathbb{R}^n$

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

c) Suppose we whiten the dataset by subtracting from each point $\mathbf{x}$ the mean μ of the dataset, and then multiplying it by $\Sigma^{-1/2}$, where Σ is the covariance of the dataset: $\mathbf{x}' = \Sigma^{-1/2}(\mathbf{x} - \mu)$. Find new parameters $\mathbf{W}', b', v'$ such that $F(\mathbf{x}, \mathbf{W}, b, v) = F(\mathbf{x}', \mathbf{W}', b', v')$.

1 pts ☐

. . . . . .

. . . . . .

d) In a neural network with n nodes, the computational cost of the backpropagation algorithm is $O(n)$. Explain why.

2 pts

. . . . . .

. . . . . .

. . . . . .

e) Describe two functions of pooling layers in a Convolutional Neural Network (CNN).

1 pts

. . . . . .

. . . . . .

f) Assume you are given a dataset containing gray-scale images of size $n \times n$. Before classifying the images you want to reduce their size to be smaller than 3×3 for which you are allowed to use consecutive 3×3 convolutional layers with stride 1 and no pooling or padding. How many such consecutive layers would you need?

2 pts

. . . . . .

. . . . . .

. . . . . .

3.2 True/False Questions

Which of the following claims are true/false? *(1 point per correct answer, -1 point per incorrect answer, non-negative total points in any case)*

5 pts

a) One advantage of using ReLUs as activation functions instead of sigmoids is that they reduce the likelihood of the gradients to vanish.

[ ] True [ ] False

b) Convolutional Neural Networks have more parameters than fully connected networks with the same number of layers and the same number of neurons in each layer.

[ ] True [ ] False

c) Neural networks can be used for regression, as well as classification.
[ ] True [ ] False

d) A max-pooling layer that reduces a 10×10 image to 5×5 has the same number of parameters as a convolutional layer with 10 3×3 filters.
[ ] True [ ] False

e) A single perceptron can compute the XOR function.
[ ] True [ ] False

3.3 Clustering

a) We are given a dataset of points $\{-2, 9, 1, -3, 6, 5, 4, 8\}$ in $\mathbb{R}$. Cluster this dataset using the K-means algorithm with $K = 2$, initialized at the two random clusters $C_1 = \{9, -2, 5, 8\}$ and $C_2 = \{6, 1, -3, 4\}$. Describe all steps carefully.

4 pts

b) Consider a set of N data points $\mathbf{X} := (x_1, \ldots, x_N), x_i \in \mathbb{R}^d$. We want to model this data using a mixture of K Gaussian distributions $\mathcal{N}(x|\mu_k, \Sigma_k)$. The GMM for one data point is defined as, $p_\theta(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$, where $\theta = \{\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K\}$ are the model parameters. Compute the variance of $p_\theta(x)$.

2 pts

c) How many parameters does this distribution have? Justify your answer.

2 pts

3.4 K-means vs. Gaussian mixture model (GMM)

The GMM and the K-means algorithm are closely related - the latter is a special case of GMM. The likelihood of a GMM with Z denoting the latent components can be typically expressed as

$$P(X) = \sum_z P(X|Z = z)P(Z = z),$$

where $P(X|Z)$ is the (multivariate) Gaussian likelihood conditioned on the mixture component and $P(Z)$ is the prior on the components. Such a likelihood formulation can also be used to describe a K-means clustering model.

a) How does the prior $P(Z)$ differ between GMM and K-means?

1 pts ☐

. . . . . .

. . . . . .

b) How does the covariance matrix in the Gaussian likelihood function differ between GMM and K-means?

2 pts ☐

. . . . . .

. . . . . .

c) Sketch (i.e. draw or describe) a dataset for $k = 2$ or $k = 3$ on which K-Means would work poorly, but a Gaussian Mixture Model with the same number of mixtures k would achieve a lower error and briefly explain why.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

d) What are two advantages of K-means in comparison to GMMs?

1 pts ☐

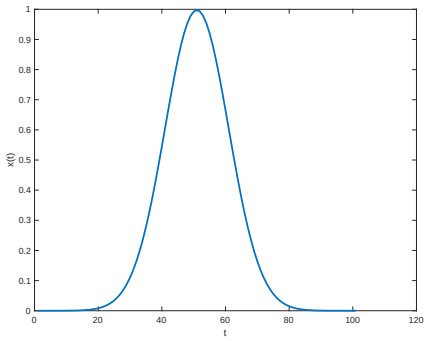
. . . . . .

. . . . . .

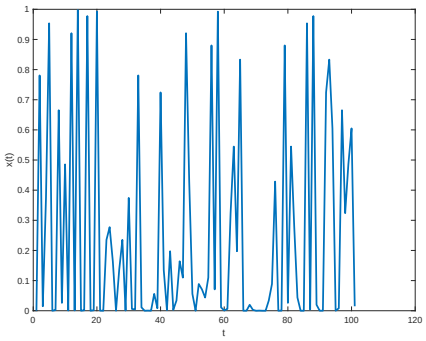
4 Sparse Coding and Dictionary Learning (30 pts)

4.1 1-D Fourier transform (6 pts)

Let $\mathbf{x} \in \mathbb{R}^d$ be a given non-sparse signal and σ is a permutation of the set of indices $\{1, \ldots, d\}$. The following two figures show $\mathbf{x}$ after applying two different permutations of the indices denoted by σ_1 and σ_2.



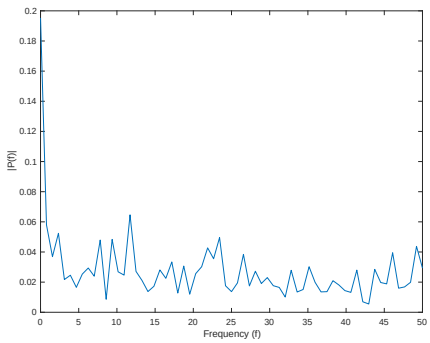
a. permutation σ_1



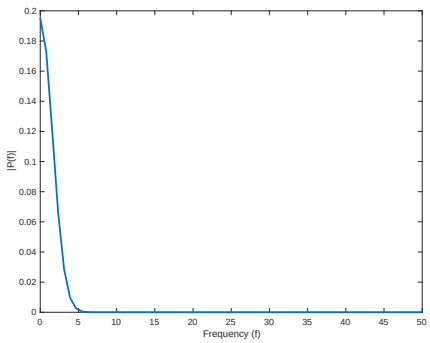
b. permutation σ_2

a) We are given the following Fourier transforms of the permuted signals. Indicate which signal gave rise to which Fourier transform?

2 pts



1. [ ]



2. [ ]

b) Suppose that we now apply a sparse coding algorithm to keep the coefficients corresponding to the low frequencies. For which signal do we get the highest reconstruction error? Justify your choice.

2 pts

. . . . . .

. . . . . .

c) Explain why we do not permute the coordinates of a vector in sparse coding, even though it might reduce the reconstruction error.

2 pts

. . . . . .

. . . . . .

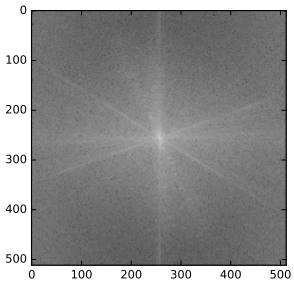
4.2 2-D Fourier transform (6 pts)

We are given the following two images.

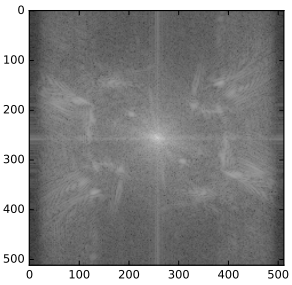
Image a

Image b

a) Below are two plots showing the magnitude of the spectrum of the 2-d cosine transform of images a and b. Brighter regions in these images correspond to larger values in the cosine



1.



2.

transform. Recall that low frequencies are located at the center of the transform images. Which cosine transform corresponds to image a? Justify your choice.

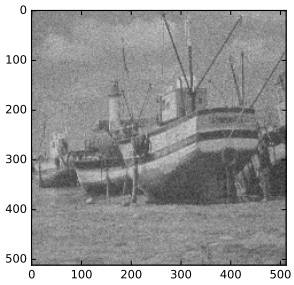
3 pts

. . . . . .

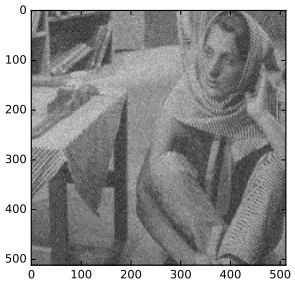
. . . . . .

b) Suppose that we want to denoise the following noisy versions of images a and b. Which noisy image can be denoised better by low-pass filtering (i.e. achieve a lower reconstruction error)? Briefly justify your choice.

3 pts



a.



b.

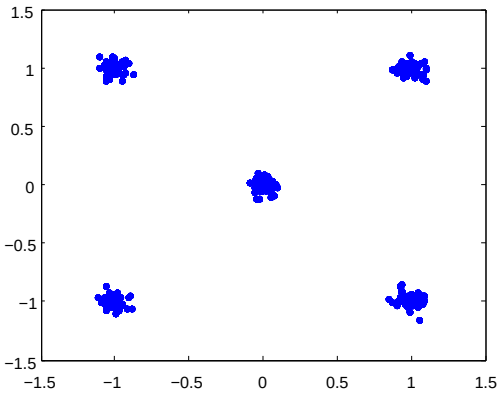
. . . . . .

. . . . . .

. . . . . .

4.3 Sparse coding with an over-complete dictionary (6 pts)

We are given the following set of 2-dimensional data points generated from a mixture of spherical Gaussians. Let $\mathbf{X}$ be a $2 \times n$ matrix whose columns contain the data points.



a) Draw a set of over-complete basis vectors on the figure such that the resulting encoding would yield a sparser representation of data points. You can ignore the variance of the data inside each cluster.

3 pts

b) We use dictionary learning to decompose the data matrix as

$$\mathbf{X} = \mathbf{U}\mathbf{Z} \tag{1}$$

where the dictionary matrix $\mathbf{U} \in \mathbb{R}^{2\times 4}$ (which means there are only 4 atoms in the dictionary) and $\mathbf{Z}$ is a sparse matrix in $\mathbb{R}^{4\times n}$. How can we approximate the centers of the mixture components using this decomposition?

3 pts

4.4 Compressed Sensing (12 pts)

We are given n gold coins. One of these coins is counterfeit, its mass being different from the other coins. Our goal is to find the counterfeit coin using a scale to make a *limited* number of measurements. We address this problem in several steps.

a) We label the coins with indices from $\{1, \ldots, n\}$. We introduce the weight vector $\mathbf{w} \in \mathbb{R}^n$ whose i-th element is the weight of the coin i. Is this vector sparse?

1 pts

b) If the weight of a gold coin is the known constant δ, is the vector $\mathbf{x} = \mathbf{w} - \delta$ sparse? Why (not)?

1 pts

c) We pick a random subset of coins denoted $\mathcal{S}$ and measure their weights. We denote the weight of the coins in $\mathcal{S}$ as y. Consider the scaling vector $\mathbf{u} \in \{0,1\}^n$ whose i-th element is 1 if the coin i is included in $\mathcal{S}$, otherwise the element is 0. What is the relation between y and the vectors $\mathbf{w}$ and $\mathbf{u}$?

2 pts ☐

. . . . . .

. . . . . .

d) We repeat the scaling operation k times (independently) using scaling vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ and we observe the weight values $y_1, \ldots, y_k$. Let the i-th row of the matrix $\mathbf{U}$ be $\mathbf{u}_i$. Similarly, the vector $\mathbf{y}$ contains y_i in its i-th coordinate. How we can obtain $\mathbf{y}$ from $\mathbf{w}$ and $\mathbf{U}$?

1 pts ☐

. . . . . .

. . . . . .

e) Write the underlying optimization problem for finding the counterfeit coin from measurements $\mathbf{y}$ and $\mathbf{U}$. [**Hint:** Use results of part b. and d.]

3 pts ☐

. . . . . .

. . . . . .

f) Write the convex relaxation of the problem formulated in question e.

2 pts ☐

. . . . . .

. . . . . .

g) Assume that the random scaling matrix $\mathbf{U}$ is similar to a random matrix with i.i.d. Gaussian elements. Based on your knowledge of compressed sensing, what is the minimum number of measurements k required to find the counterfeit coin?

2 pts ☐

. . . . . .

. . . . . .