ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computational Intelligence Lab

2018

Prof. T. Hofmann

Final Exam

August 6, 2018

First and Last name: _____

Student ID (Legi) Nr: _____

Signature: _____

General Remarks

- Please check that you have all 26 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	NMF, pLSA and Word Embedding	30		
3	Sparse Coding and Dictionary Learning	30		
4	Neural Networks and Generative Models	30		
Total		120		

Grade: .....

1 Dimensionality reduction (30 pts)

1.1 Eigendecomposition Basics (5 pts)

Suppose $\mathbf{A}$ is a real $d \times d$ matrix whose inverse $\mathbf{A}^{-1}$ exists and recall that for an eigenvalue $\lambda \in \mathbb{R}$ with corresponding normalized eigenvector $\mathbf{v} \in \mathbb{R}^d$ we have

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \text{ and } \|\mathbf{v}\|_2 = 1$$

Prove the following statements

a) If λ is an eigenvalue of $\mathbf{A}$, then λ^2 is an eigenvalue of $\mathbf{A}^2$.

1 pts ☐

. . . . . .

b) If λ is an eigenvalue of $\mathbf{A}$, then $1/\lambda$ is an eigenvalue of $\mathbf{A}^{-1}$.

1 pts ☐

. . . . . .

c) Given a vector $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} \neq \vec{0}$, the operation $(\mathbf{I}_d - \mathbf{v}\mathbf{v}^\intercal)\mathbf{x}$ results in a vector that is orthogonal to $\mathbf{v}$. ($\mathbf{I}_d = \text{diag}(1, \ldots, 1)_d$ is the d-dimensional identity matrix)

1 pts ☐

. . . . . .

d) Let $\mathbf{A}$ be a symmetric matrix with distinct, non-zero eigenvalues. Prove that two given eigenvectors $\mathbf{v}_i$ and $\mathbf{v}_j$ are orthogonal.
Hint: First show that for any two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ we have $\mathbf{x}_1^\intercal \mathbf{A}\mathbf{x}_2 = \mathbf{x}_2^\intercal \mathbf{A}\mathbf{x}_1$.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

1.2 Data whitening (5 pts)

Eigendecompositions can be used to *whiten* a data set $\{\mathbf{x}^{(i)}\}_{i=1,...,N}$, where each datapoint $\mathbf{x}^{(i)} \in \mathbb{R}^m$. To do this, we write the eigenvector equation for the data set covariance matrix $\mathbf{\Sigma}$ in the form

$$\mathbf{\Sigma}\mathbf{u}_j = \lambda_j \mathbf{u}_j \quad \rightarrow \quad \mathbf{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}, \tag{1}$$

where $\mathbf{\Lambda}$ is an $m \times m$ diagonal matrix $\mathrm{diag}(\lambda_1, ..., \lambda_m)$ and $\mathbf{U}$ is an $m \times m$ orthogonal matrix with columns given by $\mathbf{u}_j$, i.e. $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_m]$. We then define for each data point $\mathbf{x}^{(i)}$ a transformed (i.e. *whitened*) value given by

$$\mathbf{y}^{(i)} := \mathbf{\Lambda}^{-1/2}\mathbf{U}^\intercal(\mathbf{x}^{(i)} - \bar{\mathbf{x}}), \quad \forall i = 1, \ldots, N \tag{2}$$

where $\bar{\mathbf{x}} = 1/N \sum_{i=1}^{N} \mathbf{x}^{(i)}$ denotes the sample mean and $\mathbf{\Lambda}^{-1/2}$ is the diagonal matrix given by the $\lambda_j^{-1/2}$, i.e. $\mathbf{\Lambda}^{-1/2} = \mathrm{diag}(\lambda_1^{-1/2}, \ldots, \lambda_m^{-1/2})$.

a) Compute the mean of the transformed set $\{\mathbf{y}^{(i)}\}_{i=1,..,N}$.

1.5 pts

. . . . . .

. . . . . .

b) Compute the covariance matrix of the transformed set $\{\mathbf{y}^{(i)}\}_{i=1,..,N}$.

2.5 pts

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

c) What can you say about the correlations between different dimensions of the transformed data?

1 pts

. . . . . .

. . . . . .

1.3 Dimensionality Reduction (5 pts)

Consider a two-layer linear perceptron of the form shown in Fig. 1, with m inputs, m outputs and $d < m$ hidden units, where each activation function is linear. You may assume that the data is centered in 0 and that the network has no bias variables. The targets used to train the network are simply the inputs themselves, i.e. the network is trained to map each input vector onto itself. The network weight matrices $\mathbf{W}_1, \mathbf{W}_2$ are determined by minimizing the loss

$$L(\mathbf{x}^{(1:N)}; \mathbf{W}_1, \mathbf{W}_2) = \frac{1}{N} \sum_{i=1}^{N} ||\mathbf{y}(\mathbf{x}^{(i)}; \mathbf{W}_1, \mathbf{W}_2) - \mathbf{x}^{(i)}||_2^2, \tag{3}$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^m$ denotes the i-th data point ($i \in \{1, ..., N\}$) and $\mathbf{W}_j$ the weight matrix of layer $j \in \{1, 2\}$.

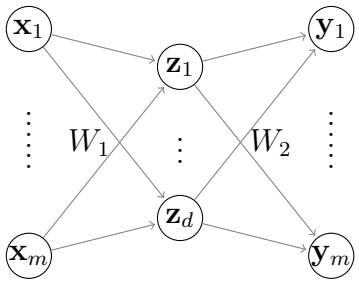


Figure 1: Two-layer linear neural network.

It can be shown that the loss function has a unique global minimum (you may assume that this holds without having to prove it) and that at this minimum, the trained network performs a dimensionality reduction followed by a reconstruction.

a) By analogy to PCA, what are the optimal weight matrices $\mathbf{W}_1, \mathbf{W}_2$?

Hint: No need to compute any derivatives.

2 pts

. . . . . .

. . . . . .

. . . . . .

b) What are the hidden units $\mathbf{z}(\mathbf{x}^{(i)}, \mathbf{W}_1)$ of the trained network as a function of the input $\mathbf{x}^{(i)}$ and optimal weights?

1 pts

. . . . . .

c) What is the optimal reconstruction $\mathbf{y}(\mathbf{x}^{(i)}; \mathbf{W}_1, \mathbf{W}_2)$ as a function of the input $\mathbf{x}^{(i)}$ and optimal weights?

1 pts

. . . . . .

d) Suggest two extensions/modifications of the neural network architecture that could improve the reconstruction error.

1 pts

. . . . . .

. . . . . .

1.4 Principal Component Analysis (9 pts)

Let $\mathbf{x} \in \mathbb{R}^m$ be a random vector of m features with symmetric positive definite covariance matrix $\mathbf{\Sigma} = \mathbf{E}(\mathbf{x}\mathbf{x}^\intercal) - \mathbf{E}(\mathbf{x})\mathbf{E}(\mathbf{x})^\intercal$. Assume the spectrum of $\mathbf{\Sigma}$ is bounded as follows

$$0 < \lambda_{\min}(\mathbf{\Sigma}) < \lambda_{\max}(\mathbf{\Sigma}) < \infty. \tag{4}$$

In PCA we observe N i.i.d. samples $\mathbf{X} := (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\intercal$ and want to do dimensionality reduction on $\mathbf{X}$ in the sense that we want to project the observations $\mathbf{x}^{(i)}$ onto a lower dimensional subspace by linearly combining the features while preserving as much variance as possible. Suppose we are given the true covariance matrix $\mathbf{\Sigma}$ and we want to find just one direction $\mathbf{u}^*$ among which $\mathbf{x}$ is most spread out in terms of the variance along this direction, i.e. $\mathrm{Var}\left[\mathbf{x}^\intercal \mathbf{u}^*\right] \geq \mathrm{Var}\left[\mathbf{x}^\intercal \mathbf{u}\right], \ \forall \mathbf{u} \in \mathbb{R}^d$. In this case PCA gives rise to the following optimization problem

$$\max_{\mathbf{u} \in \mathbb{R}^d} \mathbf{u}^\intercal \mathbf{\Sigma} \mathbf{u}, \quad s.t. \ \|\mathbf{u}\|_2 = 1. \tag{5}$$

a) Prove that the quantity maximized in (5) indeed represents the variance of $\mathbf{x}$ along the direction $\mathbf{u}$, i.e. show that

$$\mathbf{u}^\intercal \mathbf{\Sigma} \mathbf{u} = \mathrm{Var}\left[\mathbf{u}^\intercal \mathbf{x}\right].$$

Hint: recall the definition of $\mathbf{\Sigma}$ as well as the fact that for a given random vector $\mathbf{y}$ we have $\mathrm{Var}[\mathbf{y}] = \mathbf{E}(\mathbf{y}^2) - (\mathbf{E}(\mathbf{y}))^2$.

3 pts

. . . . . .

. . . . . .

. . . . . .

For a given eigenvector one can compute its corresponding eigenvalue via the so-called Rayleigh quotient:

$$r(\mathbf{u}) := \frac{\mathbf{u}^\intercal \mathbf{\Sigma} \mathbf{u}}{\mathbf{u}^\intercal \mathbf{u}}, \tag{6}$$

that is, $r(\mathbf{v}_i) = \lambda_i$.

b) Prove that the function $r(\mathbf{u})$ is upper bounded by the largest eigenvalue $\lambda_{\max}$ of $\mathbf{\Sigma}$.

Hint: Make use of the decomposition $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\intercal$ and the fact that the columns of $\mathbf{V}$ form an orthonormal basis of $\mathbb{R}^d$.

3 pts ☐

c) Give an expression for a maximizer $\mathbf{u}^*$ of (6)?

1 pts ☐

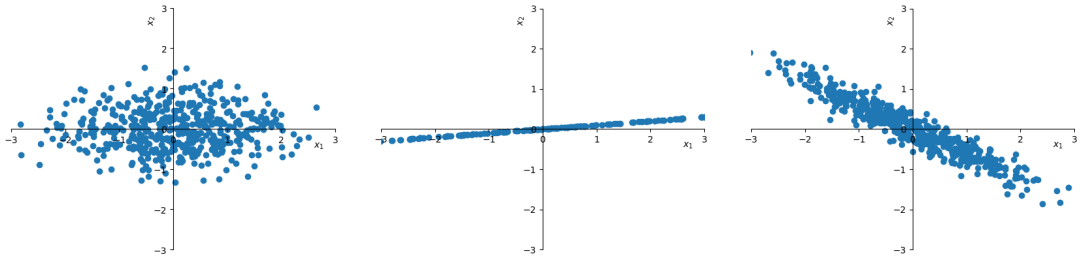
d) Show that the any eigenvector $\mathbf{v}_i$ of $\mathbf{\Sigma}$ is a first order stationary point of (6), i.e. $\nabla r(\mathbf{v}_i) = 0,\ \forall i = 1, \ldots, d$.

2 pts ☐

1.5 Data Distributions and Covariance Matrices (3 pts)

The following plots show datapoints drawn from a multivariate Gaussian distribution with zero mean and one of the below given covariance matrices $\mathbf{A} - \mathbf{F}$. Indicate which of the candidate covariance matrices corresponds to which sample $(a) - (c)$.
(1 point per correct answer, -1 point per incorrect answer, no negative total points possible)

3 pts ☐



(a) (b) (c)

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 0.5 \end{bmatrix}$$
$$\mathbf{D} = \begin{bmatrix} 5 & 0.5 \\ 0.5 & 0.05 \end{bmatrix} \quad \mathbf{E} = \begin{bmatrix} 1.25 & -0.75 \\ -0.75 & 0.5 \end{bmatrix} \quad \mathbf{F} = \begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix} \tag{7}$$

(a) .....

(b) .....

(c) .....

1.6 Multiple Choice (3 pts)

Which of the following statements are true/false? (1 point per correct answer, -1 point per incorrect answer, no negative total points possible)

3 pts ☐

a) If $\mathbf{A}$ is a real symmetric matrix with non-negative eigenvalues, then the eigenvalues and singular values of $\mathbf{A}$ coincide.

[ ] True [ ] False

b) If PCA is performed on an uncorrelated data set, the eigenvalues all equal one.

[ ] True [ ] False

c) Suppose that SVD is used for Collaborative Filtering. If two users have no common ratings, they also share no latent concepts.

[ ] True [ ] False

2 NMF, pLSA and Word Embedding (30 pts)

2.1 Non-negative Matrix Factorization (8 pts)

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \mathbf{X} \approx \mathbf{U}^\top \mathbf{V} \\ \text{s.t.} \quad & u_{zi}, v_{zj} \geq 0 \quad (\forall i, j, z). \end{aligned}$$

a) Write at least one example where non-negativity constraints are useful in matrix factorization.

1 pts ☐

. . . . . .

. . . . . .

b) Let us consider the problem of non-negative matrix factorization in the following setting,

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v}} \quad & J(\mathbf{u}, \mathbf{v}) = \frac{1}{2}\|\mathbf{X} - \mathbf{u}\mathbf{v}^\top\|_F^2 + \frac{1}{2}\lambda\|\mathbf{u}\|_F^2 + \frac{1}{2}\lambda\|\mathbf{v}\|_F^2 \\ \text{s.t.} \quad & u_{zi}, v_{zj} \geq 0 \quad (\forall i, j, z), \end{aligned}$$

where $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{v} \in \mathbb{R}^n$, $\lambda > 0$.

1. Prove that the objective function $J(\mathbf{u}, \mathbf{v})$ is convex with respect to $\mathbf{u}$.

   2 pts ☐

   . . . . . .

   . . . . . .

   . . . . . .

2. Is the objective $J(\mathbf{u}, \mathbf{v})$ a convex function (with respect to all its arguments).

   1 pts ☐

   . . . . . .

c) Let us consider a multidimensional non-negative matrix factorization problem,

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & J(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^\top\mathbf{V}\|_F^2 + \frac{1}{2}\lambda\|\mathbf{U}\|_F^2 + \frac{1}{2}\lambda\|\mathbf{V}\|_F^2 \\ \text{s.t.} \quad & u_{zi}, v_{zj} \geq 0 \quad (\forall i, j, z), \end{aligned} \tag{8}$$

where $\mathbf{X} \in \mathbb{R}^{m \times n}$, $\mathbf{U} \in \mathbb{R}^{k \times m}$, $\mathbf{V} \in \mathbb{R}^{k \times n}$, $\lambda > 0$.

1. Outline the Projected ALS algorithm for approximately solving the NMF problem (8).

2 pts ☐

2. Derive the update rule for a column $\mathbf{v}_j \in \mathbb{R}^k$ (before projecting).

2 pts ☐

2.2 pLSA, EM (8 pts)

Assume that we have a corpus containing the following 3 documents,

- 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$\}$.

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

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

1 pts ☐

. . . . . .

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

1 pts ☐

. . . . . .

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

Table 1: The co-occurrence matrix $\mathbf{X}$ of pLSA

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 in 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. Introduce the variational parameters q_{zij} and derive the variational lower bound on $\mathcal{L}(\mathbf{U}, \mathbf{V})$ using Jensen's inequality

$$\log\left(\frac{\sum_z q_z a_z}{\sum_z q_z}\right) \geq \frac{\sum_z q_z \log(a_z)}{\sum q_z}.$$

.

1 pts ☐

. . . . . .

. . . . . .

. . . . . .

4. Write down the constraints on the introduced variational parameters q_{zij}.

1 pts ☐

. . . . . .

. . . . . .

5. Write down the Lagrangian $L(\mathbf{U}, \mathbf{V}, \{q_{zij}\}, \{\lambda_{ij}\}, \dots)$ for maximization of the variational lower bound (use the result derived in 3.) and derive the update rule for q_{zij}.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

2.3 Word Embedding, Stochastic Gradient Descent (10 pts)

GloVe is a popular model to learn word embeddings based on the co-occurrence of words in the corpus. Its objective is the weighted least squares fit of log-counts

$$\mathcal{H}(\theta; \mathbf{N}) = \sum_{i,j} f(n_{ij}) \left(\log n_{ij} - \log \tilde{p}_\theta(w_i|w_j)\right)^2,$$

with unnormalized distribution

$$\tilde{p}_\theta(w_i|w_j) = \exp\left[\langle \tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_j \rangle + b_i + d_j\right],$$

where f is the weighting function. $\mathbf{x}_i := [\tilde{\mathbf{x}}_i, b_i]$ is the word embedding of word w_i, and $\mathbf{y}_i := [\tilde{\mathbf{y}}_j, d_j]$ is the context embedding of word w_j. θ encapsulates the embeddings $\mathbf{x}_i, \mathbf{y}_j$, and $\mathbf{N} = [n_{ij}]$.

a) What is the meaning of n_{ij} in the GloVe objective?

1 pts ☐

. . . . . .

. . . . . .

b) Describe the function f (write down its form or briefly draw it). What is the purpose of f (write down at least two aspects)?

3 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

c) One way to minimize the GloVe objective is using Full-Batch Gradient Descent. Derive the full gradient with respect to the embedding vectors $\mathbf{x}_i$ and $\mathbf{y}_j$. Explain why one usually cannot use Full-Batch Gradient Descent in practice?

4 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

d) One alternative to Full-Batch Gradient Descent is the Stochastic Gradient Descent. Briefly explain how it works.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

2.4 Multiple Choice (4 pts)

Which of the following statements are true/false? (1 point per correct answer, -1 point per incorrect answer, no negative total points possible)

4 pts ☐

1. "One-hot" word vector representations usually cannot capture similarity of words.

   [ ] True [ ] False

2. The GloVe objective uses unnormalized models, which do not need computation of partition function. It uses a two-sided loss function.

   [ ] True [ ] False

3. The GloVe model introduces context vectors to increase modeling flexibility, which makes the model dimensionality larger at the same time.

   [ ] True [ ] False

4. The GloVe model solves a matrix completion problem.

   [ ] True [ ] False

3 Sparse Coding and Dictionary Learning (30 pts)

3.1 True/False Questions (6 pts)

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)*

a) Every two atoms in a dictionary must be orthogonal.

[ ] True [ ] False

b) We only need to constrain on the sparsity of the representation in dictionary learning.

[ ] True [ ] False

c) In dictionary learning, the objective is not jointly convex in the dictionary and the sparse representations.

[ ] True [ ] False

d) The advantage of PCA over a fix overcomplete basis for sparse coding is that we only need to transmit the eigenvalues and not the whole basis.

[ ] True [ ] False

e) Adding atoms to a fixed overcomplete atom set leads to sparser coding.

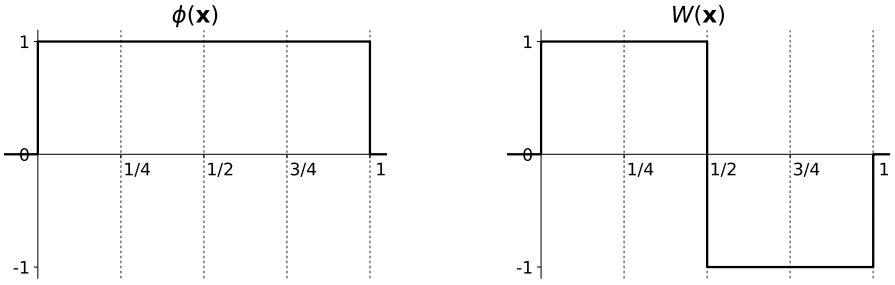
[ ] True [ ] False

f) Adding atoms to a fixed overcomplete atom set increases its coherence.

[ ] True [ ] False

3.2 Haar transform and sparse coding (6 pts)

We are given two basis vectors of a discrete Haar wavelet transform, one for the scaling function $\phi(x) = (1, 1, 1, 1)$ and the other for a mother Haar wavelet $W(x) = (1, 1, -1, -1)$. These two basis vectors are shown below.



a) Write down the basis vectors for the dilated Haar wavelet function $W(2x)$ and translated dilated Haar wavelet function $W(2x-1)$.

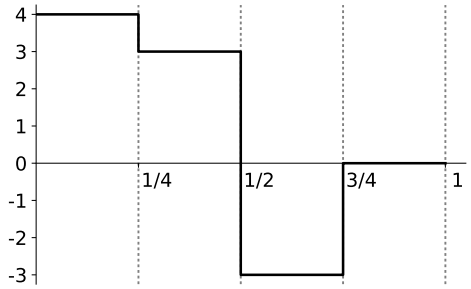
2 pts

. . . . . .

. . . . . .

b) Decompose the following signal $s(x) = (4, -2, 0, 4)$ with respect to the Haar wavelets using the equation in the box below. [Note that the coefficients can be negative.]

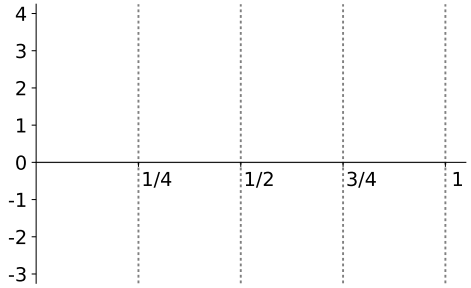
2 pts



$$s(x) = [\quad] \times \phi(x) + [\quad] \times W(x) + [\quad] \times W(2x) + [\quad] \times W(2x-1)$$

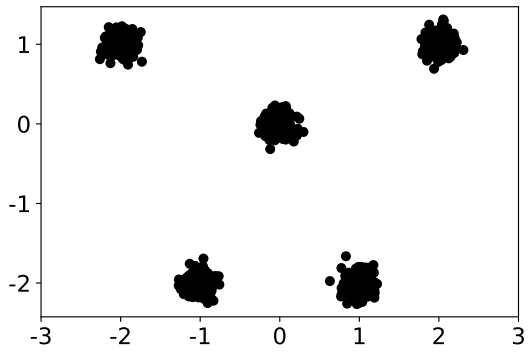
c) If we want to use only three wavelet basis vectors to reconstruct the above signal with the minimum reconstruction error, how would the reconstructed signal look like? Plot the reconstructed signal using the following coordinate system.

2 pts



3.3 Dictionary learning (8 pts)

We are given a set of 2-dimensional data generated from a mixture of Gaussians, as shown in the following figure. Let $\mathbf{X} \in \mathbb{R}^{2\times N}$ denote the dataset, where a column of $\mathbf{X}$ is a data point.



a) Write down the dictionary learning objective with constraint(s) in terms of the dictionary $\mathbf{U}$ and sparse representations $\mathbf{Z}$.

3 pts

. . . . . .

. . . . . .

b) We can use an overcomplete dictionary $\mathbf{U} \in \mathbb{R}^{2\times 4}$ with only 4 atoms to learn a sparser representation of the data points with mininum reconstruction errors. Write down the atoms of the dictionary.

2 pts

$$\mathbf{u}_0 = (\quad,\quad)^T$$
$$\mathbf{u}_1 = (\quad,\quad)^T$$
$$\mathbf{u}_2 = (\quad,\quad)^T$$
$$\mathbf{u}_3 = (\quad,\quad)^T$$

c) Write down the sparse representations of the data clusters given the overcomplete dictionary $\mathbf{U}$ from a).

3 pts

. . . . . .

3.4 Matching Pursuit (10 pts)

Recall the MP algorithm

1: $\mathbf{z} \leftarrow \mathbf{0}$, $\mathbf{r}_0 \leftarrow \mathbf{x}$
2: **while** $\|\mathbf{z}\|_0 < K$ **do**
3: Select atom with maximum absolute correlation to residual:

$$d^* \leftarrow \underset{d}{\operatorname{argmax}} \left|\mathbf{u}_d^\top \mathbf{r}_t\right|$$

4: Update coefficient vector and residual:

$$z_{d^*} \leftarrow z_{d^*} + \mathbf{u}_{d^*}^\top \mathbf{r}_t$$

$$\mathbf{r}_{t+1} \leftarrow \mathbf{r}_t - \left(\mathbf{u}_{d^*}^\top \mathbf{r}_t\right) \mathbf{u}_{d^*}$$

5: **end while**

a) Prove that the step (3) in the algorithm minimizes the norm of the residual at each step.

2 pts

b) Prove that $\|\mathbf{r}_{t+1}\|^2 \leq \|\mathbf{r}_t\|^2$.

2 pts

c) Let $I_{\mathbf{U}} := \inf_{\mathbf{r}\in\mathbb{R}^n\setminus 0} \sup_{\mathbf{u}_d\in\mathbf{U}} \frac{\langle \mathbf{r}, \mathbf{u}_d\rangle^2}{\|\mathbf{r}\|^2}$. Note that $I_{\mathbf{U}} \in (0, 1]$. Compute $I_{\mathbf{U}}$ when $\mathbf{U}$ is the $L1$ ball in 2 and in n dimensions.

2 pts

. . . . . .

d) Using the definition of $I_{\mathbf{U}}$ prove that the norm of the residual converges to zero with the following rate: $\|\mathbf{r}_{t+1}\|^2 \leq (1 - I_{\mathbf{U}})^t \|\mathbf{r}_0\|^2$.

2 pts

. . . . . .

. . . . . .

. . . . . .

e) Assume that performing step (3) of the algorithm is too expensive, therefore, we can only afford an approximate solution to the maximization problem. Assume that the approximate solution d_t is such that $\left|\mathbf{u}_{d_t}^\top \mathbf{r}_t\right|^2 \geq \alpha \operatorname{argmax}_d \left|\mathbf{u}_d^\top \mathbf{r}_t\right|^2 = \alpha \left|\mathbf{u}_{d^*}^\top \mathbf{r}_t\right|^2$ for some fixed value $\alpha > 0$ independent of t. Prove that the norm of the residual converges to zero with the following rate: $\|\mathbf{r}_{t+1}\|^2 \leq (1 - \alpha I_{\mathbf{U}})^t \|\mathbf{r}_0\|^2$

2 pts

. . . . . .

. . . . . .

. . . . . .

4 Neural networks / Generative Models (30 pts)

4.1 Neural networks (12 points)

1. The neuron is the basic computational unit of a network, from which all architectures are composed. Draw a schematic of a neuron, clearly mark the inputs/outputs and write its transfer function.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

2. Consider a deep classification network that transforms an input $\mathbf{x} \in \mathbb{R}^d$, e.g. features, to outputs $\mathbf{y} \in \{0, 1\}$. The network has 1 input layer, 1 output layer and L hidden layers, and all layers are fully connected. Write down the transfer function of the network which maps $\mathbf{x} \mapsto \mathbf{y}$. Hereby specify the activation function to use, or use the generic symbol σ if any activation function could be used.

3 pts ☐

. . . . . .

. . . . . .

3. From the perspective of learning theory, a trainable neural network N induces a class of hypotheses $\mathcal{H}_N$, i.e. a set of functions $h : \mathbf{x} \mapsto \mathbf{y}$ mapping network inputs to network outputs. For example, consider two multilayer perceptrons N_k, N_l with k and l hidden layers, respectively, where $k > l > 0$. Each layer uses the activation function $\sigma(x) = x$. Write explicitly the hypothesis classes induced by the two networks, which you should denote by $\mathcal{H}_k$ and $\mathcal{H}_l$, and show that $\mathcal{H}_k = \mathcal{H}_l$. (You may assume that $\mathbf{x} \in \mathbb{R}^d$ and $l > d$).

4 pts ☐

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

4. Sketch the activation function `tanh` on the interval $[-2, 2]$. Compute the derivative $\sigma'(x)$of $\sigma(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and express it in terms of $\sigma(x)$.

3 pts

. . . . . .

. . . . . .

. . . . . .

. . . . . .

4.2 Convolutional Neural Networks (12 points)

1. An alternative to using CNNs is using an MLP with flattened image input. Briefly explain how this choice *would not* lead to weight sharing and *would not* lead to translation invariance.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

2. Consider images of size 32x32 with 3 channels. Compute the number of parameters of a fully connected network with one hidden layer having 32 neurons and an output layer with a neuron. The activation function is Sigmoid.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

3. For an image of the same size, compute the number of the parameters of a convolution layer with the number of kernels 8 and filter size 5x5 and stride 1 and zero-padding 1. The activation function is ReLU.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

4. In this question, we are considering a convolutional neural network designed to process 3D images, which could arise e.g. from MRI scans of the human brain. The image consists of 1 grey-scale channel, and we use 1 3D 3x3x3 filter that operates on volume cubes. Our network consists of 1 convolutional layer with a Leaky rectified linear unit defined by $f(x) = 0.01x$ if $x < 0$ and $f(x) = x$ if $x \geq 0$, which is followed by a 3×3 2D average pooling layer with a stride of 1. In contrast to the filters, the pooling is applied separatedly to each z-plane. Before computing the filters, zero-padding is used, such that

the convolved image retains the dimension $3 \times 3 \times 3$. For average pooling *no zero padding* is used, the average is computed only over present entries in the neighbourhood.

Assume you are given the following input volume, whose 3 z-planes are given by

$$I_0 = \begin{pmatrix} 0 & 2 & 3 \\ 1 & 0 & 0 \\ 0 & 0 & 5 \end{pmatrix}, I_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, I_2 = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

The sole 3D filter is similarly defined by

$$K_0 = K_2 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & 0 \end{pmatrix}, K_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 5 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Please compute the 2 entries of the convolved 3D image $C := f(I * K) \in \mathbb{R}^{3\times3\times3}$, and the 1D output $O :=$ AveragePooling$(C) \in \mathbb{R}^{1\times3}$, *only for* the coordinates [1,1,1] (the center of the cube), and the coordinates [0,0,0] (the entry in the front-bottom-left of the cube).

6 pts

4.3 Generative Models (6 points)

1. A typical technique to generate new samples from a distribution is to use a Variational Autoencoder (VAE). A VAE can be seen as a deep architecture composed of an encoder and a decoder where some some randomness is added to the encoded latent variable. Please explain what the encoder and the decoder represent in a Bayesian framework setting. Write your answer in terms of the input x, the latent variable z, the parameters of the decoder θ and the parameters of the decoder ϕ.

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

2. Let us assume that the encoder distribution is Gaussian. We can parametrize it by its mean μ and its variance σ, which are going to be the two outputs of the encoder. The input of the decoder should be a sample of this distribution, i.e. $z \sim \mathcal{N}(\mu, \sigma)$. By drawing randomly a sample, we are able to perform forward pass, however using this setting we cannot back-propagate the error through the encoder. Can you explain why?

2 pts ☐

. . . . . .

. . . . . .

. . . . . .

3. In order to solve this backpropagation issue, the reparameterization trick is applied. Please write down the reparametrization trick for a univariate Gaussian distribution?

2 pts ☐

. . . . . .

. . . . . .

. . . . . .