ECE 685D HW4

Submission Instructions:

1. Upload your Jupyter Notebook (.ipynb file).
2. Export all outputs and necessary derivations as a PDF or HTML (preferred) and upload them.

LLM policy:

The use of large language models (LLMs) is allowed for this assignment, but if you use them, you **must disclose how.**

1 Problem 1: Sparse Encoding for Denoising (25 pts)

Consider an auto-encoding like this:

$$\min_D \frac{1}{T}\sum_{t=1}^{T} \min_{h^{(t)}} \frac{1}{2}\|x^{(t)} - Dh^{(t)}\|_2^2 + \lambda\|h^{(t)}\|_1.$$

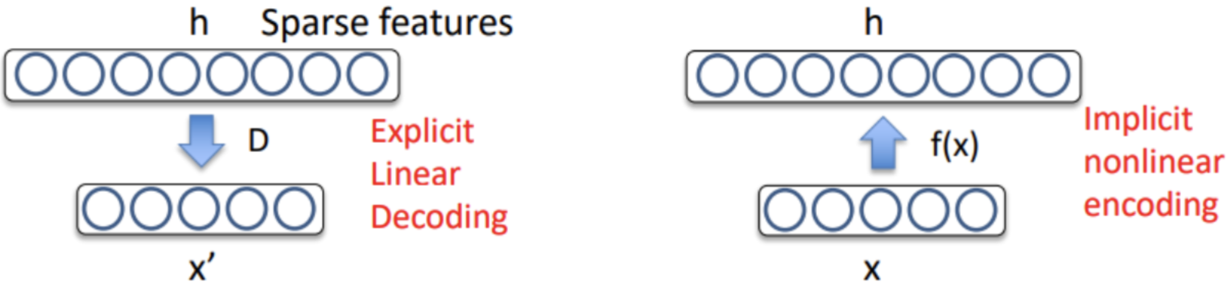

Figure 1: Over-complete auto-encoder with non-linear encoding and linear decoding modules. LASSO (L_1) regularization is used to enforce meaningful feature learning.

We will use MNIST dataset. Let the dimension of h be 1.5 times that of your input and $f(x)$ defined by your own conjecture. Let X be a batch sampled from MNIST, taking $X + \mathcal{N}(0, \sigma^2 I)$ (σ^2 of your choice, and you can clip the pixrls of noisy images between 0 and 1 to stabilize training) as the model input, and our goal is to obtain its output $\widehat{X} \approx X$. Use the default training split for your auto-encoder. You should plot

(a) 5 input-output pairs $(X + \mathcal{N}(0, \sigma^2 I), \widehat{X})$ using the data in the test set.

(b) The top 5 dictionary vectors in D whose corresponding intensity $|h_i|$ are the largest.

(c) Plot the mean square error (MSE) between noiseless image X and predicted image $\widehat{X}$ as a function of the noise level σ^2. You can do this for about 5 values of σ^2 between 0 and 1.

Hint: Using MLP for this may make the visualization easier.

2 Problem 2: Probabilistic PCA (35pts)

Let $x_1, \cdots, x_N \in \mathbb{R}^d$ be N independent observations from

$$x_j = \underbrace{Wz_j + \mu}_{\text{signal}} + \underbrace{\epsilon_j}_{\text{noise}}, \quad \text{where } \begin{bmatrix} z_j \\ \epsilon_j \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \begin{bmatrix} I_q & 0 \\ 0 & \sigma^2 I_d \end{bmatrix}\right).$$

That is, (x_j) are from the Gaussian distribution $\mathcal{N}(\mu, C)$, where the covariance matrix is $C = WW^\top + \sigma^2 I_d$. We target to estimate the parameters

$$\mu \in \mathbb{R}^d, \quad W \in \mathbb{R}^{d\times q}, \quad \text{and} \quad \sigma \in (0, \infty)$$

by maximum likelihood estimation (MLE). Here the dimension $q < \min\{d, N\}$ is known.

(a) (5pts) Write the likelihood function $\mathcal{L}(\mu, W, \sigma^2)$ given $x_1, \cdots, x_N$, and derive the MLE $\widehat{\mu}$ of μ.

(b) (10pts) Define the sample covariance matrix of $x_1, \cdots, x_N$ as

$$S = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^\top,$$

where $\overline{x} = N^{-1}\sum_{i=1}^N x_i \in \mathbb{R}^d$ is the mean of $x_1, \cdots, x_N$. Prove that $\mathcal{L}(\mu, W, \sigma^2)$ is minimized only if

$$C^{-1}W = C^{-1}SC^{-1}W.$$

Remark. You may use the following fact: for any positive definite matrices $A, X \in \mathbb{R}^{d\times d}$,

$$\frac{\mathrm{d}}{\mathrm{d}X}\log\det(X) = X^{-1}, \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}X}\mathrm{Tr}(AX^{-1}) = -X^{-1}AX^{-1}.$$

(c) (10pts) Assume W is a minimizer, and $W = QDV^\top$ the compact SVD of W, i.e. $Q \in \mathbb{R}^{d\times q}, V \in \mathbb{R}^{q\times q}$ are two column orthogonal matrices, and $D \in \mathbb{R}^{q\times q}$ is a nonnegative diagonal matrix. Show that

$$SQ = Q(D^2 + \sigma^2 I_q).$$

Using this formula to argue that all potential minimizer W has the form of

$$W = U_q\left(\Lambda_q - \sigma^2 I_q\right)^{1/2} R,$$

where the columns of U_q are q distinct eigenvectors of S, Λ is a diagonal matrix with corresponding eigenvalues and $R \in \mathbb{R}^{q\times q}$ is any orthogonal matrix.

(d) (10pts) Show that the MLE of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j(S), \tag{2.1}$$

where $\lambda_1(S) \geq \lambda_2(S) \geq \cdots \geq \lambda_d(S)$ are eigenvalues of S, including repetitions according to algebraic multiplicity and sorted in decreasing order. Derive the MLE $\widehat{W}$ of W using (c).

3 Problem 3: Gaussian-Bernoulli Restricted Boltzmann Machines (40pts)

Let the energy function of Gaussian-Bernoulli RBM take the form

$$E(v,h) = -\left(\sum_{i,j} W_{ij} h_j \frac{v_i}{\sigma_i} - \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} + \sum_j \alpha_j h_j \right).$$

(a) Use the definition $p(v,h) = e^{-E(x,h)}/Z$ to derive $p(v_i = x|h)$ and $p(h_j = 1|v)$. You may leave $p(v_i = x|h)$ in an integral form once you cancel the terms involving $\sum_j a_j h_j$.

(b) Train a Gaussian-Bernouilli RBM over the Fashion MNIST dataset using contrastive divergence minimization (see the lecture notes). Use the standard training and testing split available in Pytorch. Use learning rate 0.001 and batch size 128. Train the model over 25 epochs for $M = \{10, 50, 100, 250\}$, where M is the dimension of the hidden weights W and report the mean squared reconstruction error for the test dataset. A template container Bernoulli-Bernoulli has been provided to you. You may use it as a start.