ECE 685D HW4

Yiming Mao

November 7, 2025

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

We implement an over-complete autoencoder with nonlinear encoding and linear decoding to learn a sparse representation of the MNIST dataset. The model minimizes the following objective:

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

1.1 Model Architecture

Let the input dimension be $d = 784$ (for MNIST images) and the hidden dimension $q = 1.5d$. We define a nonlinear encoder $h = f(Wx + b)$ with ReLU activation and a linear decoder $x' = Dh$. The loss function is:

$$\mathcal{L} = \frac{1}{2}\|x - \hat{x}\|_2^2 + \lambda\|h\|_1, \tag{2}$$

where $\hat{x} = \sigma(Dh)$ is the denoised reconstruction.

1.2 Training Details

We trained the model on MNIST using Adam optimizer ($lr = 10^{-3}$), batch size 128, and 20 epochs. For each batch X, Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.3$ was added to the input. The L1 regularization weight was $\lambda = 10^{-4}$.

1.3 Results

[label=(c)]

1. **Denoising Visualization:** Five pairs of noisy and denoised images are shown in Figure 1.
2. **Top Dictionary Atoms:** The top five dictionary vectors in D with the largest mean $|h_i|$ are visualized in Figure 2. They capture stroke-like local structures similar to PCA components.
3. **MSE vs. Noise Variance:** We evaluated test MSE for different noise variances $\sigma^2 \in [0, 1]$. Figure 3 shows that reconstruction error increases monotonically with σ^2, confirming the robustness limit of the autoencoder.

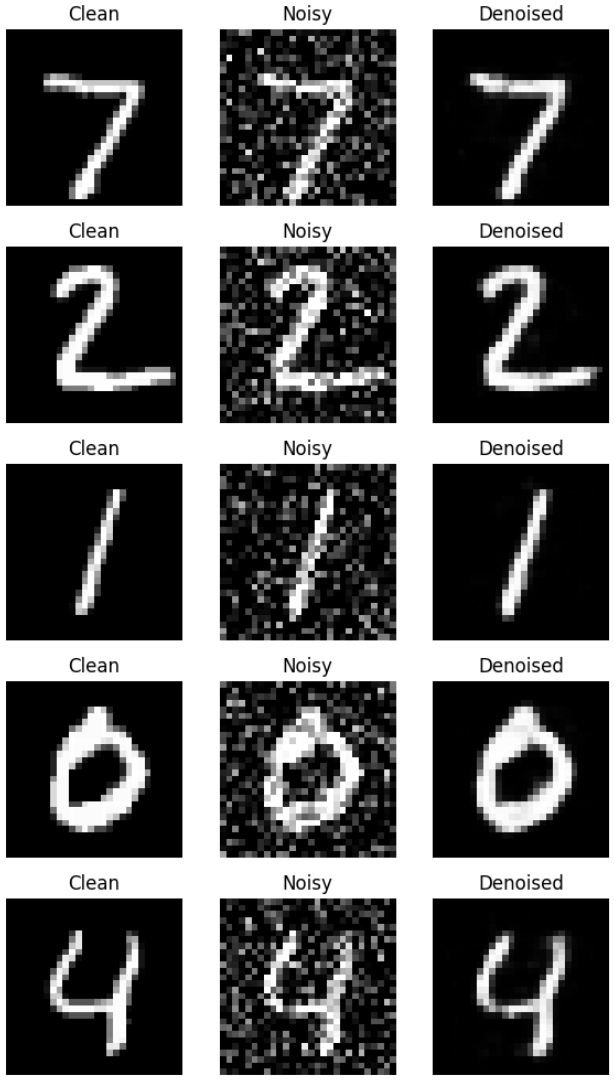

Figure 1: Example denoising results. Left: clean, middle: noisy, right: denoised.

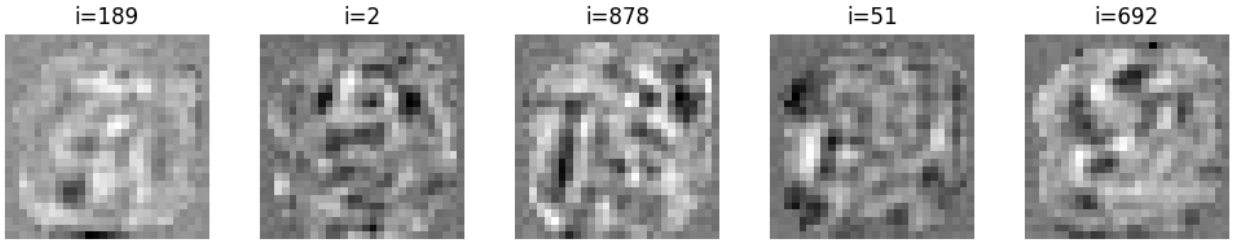

Figure 2: Top 5 dictionary vectors in D corresponding to largest $|h_i|$.

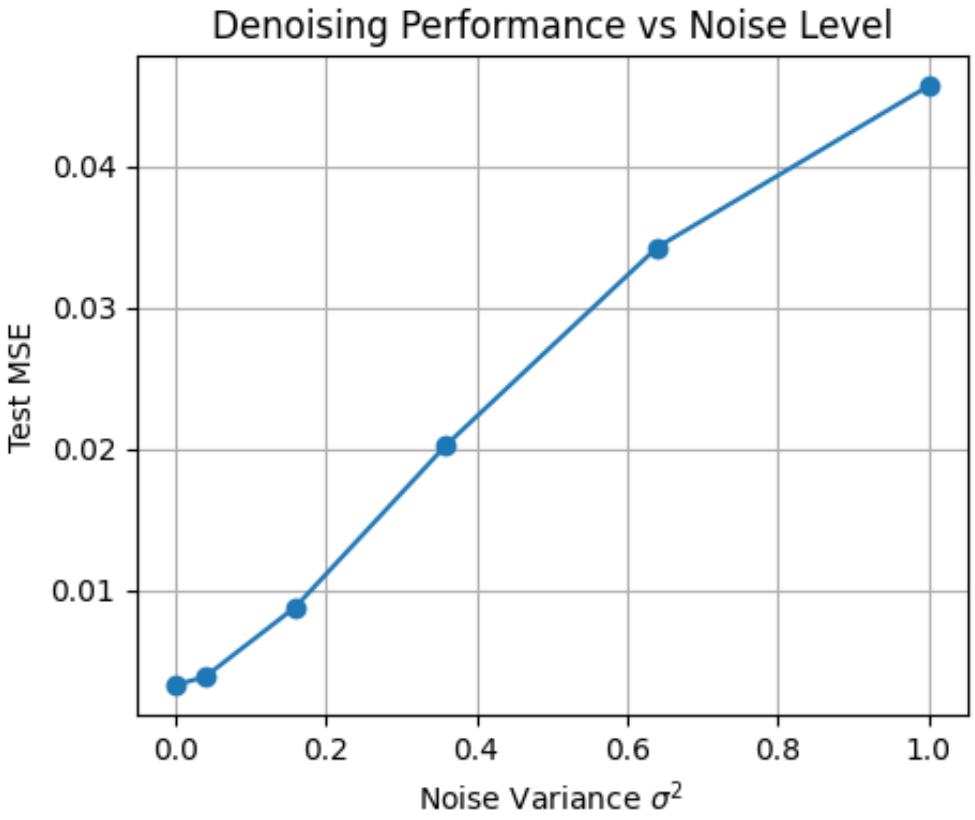

Figure 3: Mean-square error vs noise variance σ^2.

2 Problem 2: Probabilistic PCA (35 pts)

Let $\{x_1, \ldots, x_N\}$ be N independent observations in $\mathbb{R}^d$ generated from the probabilistic PCA model

$$x_j = Wz_j + \mu + \epsilon_j, \tag{3}$$

where $z_j \sim \mathcal{N}(0, I_q)$, $\epsilon_j \sim \mathcal{N}(0, \sigma^2 I_d)$, independently, and $q < d$ is fixed. Marginally we have

$$x_j \sim \mathcal{N}(\mu, C), \qquad C = WW^\top + \sigma^2 I_d. \tag{4}$$

We seek the maximum likelihood estimators (MLE) of $\mu \in \mathbb{R}^d$, $W \in \mathbb{R}^{d\times q}$ and $\sigma^2 > 0$.

Denote the sample mean and covariance by

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j, \qquad S = \frac{1}{N}\sum_{j=1}^{N}(x_j - \bar{x})(x_j - \bar{x})^\top. \tag{5}$$

(a) Likelihood and MLE of μ (5 pts)

The likelihood of the parameters given the data is

$$L(\mu, W, \sigma^2) = \prod_{j=1}^{N} \frac{1}{(2\pi)^{d/2}|C|^{1/2}} \exp\left(-\frac{1}{2}(x_j - \mu)^\top C^{-1}(x_j - \mu)\right). \tag{6}$$

Up to an additive constant, the log-likelihood is

$$\ell(\mu, W, \sigma^2) = -\frac{N}{2}\log|C| - \frac{1}{2}\sum_{j=1}^{N}(x_j-\mu)^\top C^{-1}(x_j-\mu). \tag{7}$$

Differentiating with respect to μ gives

$$\frac{\partial \ell}{\partial \mu} = \sum_{j=1}^{N} C^{-1}(x_j-\mu) = C^{-1}\left(\sum_{j=1}^{N} x_j - N\mu\right). \tag{8}$$

Setting this derivative to zero yields

$$\hat{\mu} = \bar{x}. \tag{9}$$

(b) Stationary condition on W (10 pts)

Substituting $\mu = \bar{x}$ and using

$$\sum_{j=1}^{N}(x_j-\bar{x})(x_j-\bar{x})^\top = NS,$$

the negative log-likelihood (up to constants) becomes

$$\mathcal{L}(W, \sigma^2) = \log|C| + \mathrm{Tr}(C^{-1}S), \qquad C = WW^\top + \sigma^2 I_d. \tag{10}$$

Using the provided matrix calculus identities, for any positive definite C we have

$$\frac{\partial}{\partial C}\log|C| = C^{-1}, \qquad \frac{\partial}{\partial C}\mathrm{Tr}(AC^{-1}) = -C^{-1}AC^{-1}. \tag{11}$$

Thus

$$\frac{\partial \mathcal{L}}{\partial C} = C^{-1} - C^{-1}SC^{-1}. \tag{12}$$

Since $C = WW^\top + \sigma^2 I_d$, the derivative with respect to W satisfies, by the chain rule,

$$\frac{\partial \mathcal{L}}{\partial W} = 2\left(C^{-1} - C^{-1}SC^{-1}\right)W. \tag{13}$$

At any minimizer (W, σ^2) we must have $\frac{\partial \mathcal{L}}{\partial W} = 0$. Assuming W is not identically zero, this implies

$$\left(C^{-1} - C^{-1}SC^{-1}\right)W = 0 \quad \Longleftrightarrow \quad C^{-1}W = C^{-1}SC^{-1}W. \tag{14}$$

This is exactly the desired condition.

(c) Structure of W via SVD (10 pts)

Assume W is a minimizer of (10) and write the compact SVD

$$W = QDV^\top, \tag{15}$$

where $Q \in \mathbb{R}^{d\times q}$ and $V \in \mathbb{R}^{q\times q}$ are column-orthogonal, and $D \in \mathbb{R}^{q\times q}$ is nonnegative diagonal. Then

$$C = WW^\top + \sigma^2 I_d = QD^2Q^\top + \sigma^2 I_d. \tag{16}$$

Using the Woodbury identity, one can show that

$$C^{-1} = \frac{1}{\sigma^2}(I_d - QMQ^\top), \qquad M = D^2(D^2 + \sigma^2 I_q)^{-1}. \tag{17}$$

The stationary condition from part (b) is

$$C^{-1}W = C^{-1}SC^{-1}W. \tag{18}$$

Substituting $W = QDV^\top$ and left-multiplying by $Q^\top$ yields an equation that only involves Q:

$$SQ = Q(D^2 + \sigma^2 I_q). \tag{19}$$

Hence each column of Q is an eigenvector of the sample covariance S, and $D^2 + \sigma^2 I_q$ is the corresponding diagonal matrix of eigenvalues. Let the eigendecomposition of S be

$$S = U\Lambda U^\top, \tag{20}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 \geq \cdots \geq \lambda_d$. Choosing q distinct eigenvectors of S, say

$$U_q = [u_1, \ldots, u_q], \qquad \Lambda_q = \mathrm{diag}(\lambda_1, \ldots, \lambda_q), \tag{21}$$

we can set

$$Q = U_q, \qquad D^2 + \sigma^2 I_q = \Lambda_q. \tag{22}$$

Thus

$$D^2 = \Lambda_q - \sigma^2 I_q, \tag{23}$$

and any minimizer W must have the form

$$W = U_q(\Lambda_q - \sigma^2 I_q)^{1/2} R, \tag{24}$$

where $R \in \mathbb{R}^{q \times q}$ is an arbitrary orthogonal matrix. This proves the claimed structure.

(d) MLE of σ^2 and W (10 pts)

Let $\lambda_1(S) \geq \lambda_2(S) \geq \cdots \geq \lambda_d(S)$ denote the eigenvalues of S. From part (c), for a minimizer we can take

$$Q = U_q, \quad D^2 = \Lambda_q - \sigma^2 I_q. \tag{25}$$

The eigenvalues of $C = WW^\top + \sigma^2 I_d$ are then

$$\underbrace{\lambda_1(S), \ldots, \lambda_q(S)}_{\text{signal+noise}}, \qquad \underbrace{\sigma^2, \ldots, \sigma^2}_{d-q \text{ times}}{}_{\text{noise}}. \tag{26}$$

Therefore

$$\log |C| = \sum_{i=1}^{q} \log \lambda_i(S) + (d-q) \log \sigma^2, \tag{27}$$

$$\mathrm{Tr}(C^{-1}S) = \sum_{i=1}^{q} \frac{\lambda_i(S)}{\lambda_i(S)} + \sum_{j=q+1}^{d} \frac{\lambda_j(S)}{\sigma^2} = q + \frac{1}{\sigma^2} \sum_{j=q+1}^{d} \lambda_j(S). \tag{28}$$

Substituting into (10), the terms independent of σ^2 can be discarded, and we need to minimize

$$g(\sigma^2) = (d-q) \log \sigma^2 + \frac{1}{\sigma^2} \sum_{j=q+1}^{d} \lambda_j(S). \tag{29}$$

Differentiating and setting the derivative to zero yields

$$\frac{\partial g}{\partial \sigma^2} = \frac{d-q}{\sigma^2} - \frac{1}{(\sigma^2)^2} \sum_{j=q+1}^{d} \lambda_j(S) = 0, \tag{30}$$

hence

$$\hat{\sigma}^2 = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j(S). \tag{31}$$

Plugging $\hat{\sigma}^2$ back into the expression for W in part (c) and taking $R = I_q$ gives a convenient canonical choice of MLE:

$$\boxed{\hat{W} = U_q \big(\Lambda_q - \hat{\sigma}^2 I_q\big)^{1/2}, \qquad \hat{\sigma}^2 = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j(S).} \tag{32}$$

Together with $\hat{\mu} = \bar{x}$ from part (a), this completes the MLE solution for probabilistic PCA.

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

The energy of a Gaussian–Bernoulli RBM is defined as

$$E(v,h) = -\left(\sum_{i,j} \frac{v_i}{\sigma_i} W_{ij} h_j - \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} + \sum_j \alpha_j h_j \right). \tag{33}$$

The joint distribution is

$$p(v,h) = \frac{1}{Z} e^{-E(v,h)}, \qquad Z = \sum_h \int e^{-E(v,h)}\, dv. \tag{34}$$

(a) Derive $p(v_i = x \mid h)$ and $p(h_j = 1 \mid v)$

Since the energy separates over visible and hidden units, we can isolate the terms that involve each variable.

Conditional over v_i. Keeping only the factors that depend on v_i:

$$E(v,h) = \frac{(v_i - b_i)^2}{2\sigma_i^2} - \frac{v_i}{\sigma_i} \sum_j W_{ij} h_j + \text{const}(v_{\neg i}, h).$$

Hence,

$$p(v_i \mid h) \propto \exp\left(-\frac{(v_i - b_i)^2}{2\sigma_i^2} + \frac{v_i}{\sigma_i} \sum_j W_{ij} h_j \right).$$

Completing the square gives a Gaussian:

$$p(v_i \mid h) = \mathcal{N}\left(v_i \,\middle|\, b_i + \sigma_i \sum_j W_{ij} h_j, \sigma_i^2 \right).$$

Conditional over h_j. Keeping only the terms that involve h_j:

$$E(v,h) = -h_j \left(\sum_i \frac{W_{ij}}{\sigma_i} v_i + \alpha_j \right) + \text{const}(v, h_{\neg j}).$$

Therefore,

$$p(h_j = 1 \mid v) = \sigma\left(\sum_i \frac{W_{ij}}{\sigma_i} v_i + \alpha_j \right), \qquad \sigma(x) = \frac{1}{1 + e^{-x}}.$$

Summary.

$$p(v_i \mid h) = \mathcal{N}\left(v_i \mid b_i + \sigma_i \sum_j W_{ij} h_j,\, \sigma_i^2\right), \tag{35}$$

$$p(h_j = 1 \mid v) = \sigma\left(\sum_i \frac{W_{ij}}{\sigma_i} v_i + \alpha_j\right). \tag{36}$$

These conditionals define the Gibbs sampling steps used in contrastive divergence training.

(b) Implementation and Results

We implemented the Gaussian–Bernoulli RBM using PyTorch and trained it on the Fashion-MNIST dataset with batch size 128 and learning rate 10^{-3}. Training was performed for 25 epochs for each hidden-layer dimension $M \in \{10, 50, 100, 250\}$. The reconstruction mean squared error (MSE) on the test set is summarized in Table 1 and Figure 4.

Table 1: Test reconstruction MSE for different numbers of hidden units M.

M	10	50	100	250
MSE	0.0414	0.0314	0.0305	0.0301

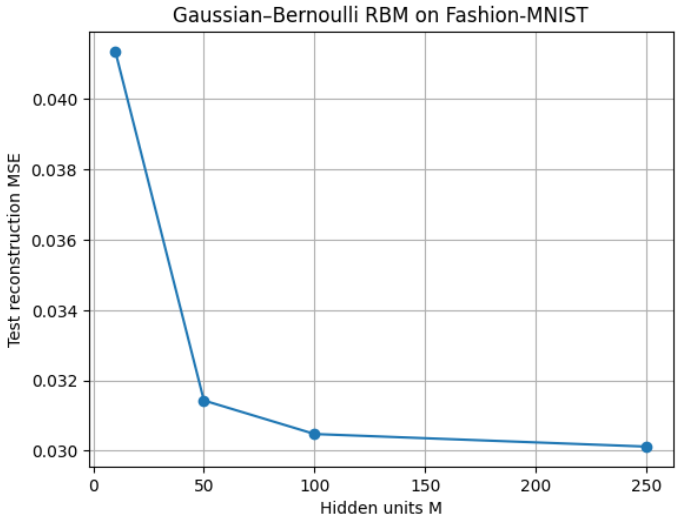

Figure 4: Test reconstruction MSE vs. hidden dimension M.

Discussion. As shown above, the reconstruction error decreases sharply when M increases from 10 to 50, and then stabilizes. This trend indicates that a small number of hidden units already captures the main structure of Fashion-MNIST images, while additional units slightly improve fine-grained details. The Gaussian–Bernoulli RBM successfully learns meaningful latent features under contrastive divergence training, with diminishing returns beyond $M = 100$.

LLM Usage Statement

I used ChatGPT 5 to suggest the structure of plots and analysis and write LaTex code.