ECE 685D HW1

Yiming Mao

September 19, 2025

Problem 1: Linear Regression on the Concrete Strength Dataset

Q1.1 Closed-form Solution

We implemented a multiple linear regression model on the Concrete Strength dataset. The model is defined as

$$Y = X\hat{\beta} + \hat{\epsilon}, \qquad \hat{\epsilon} \sim \mathcal{N}(0, \sigma^2 I),$$

where $X \in \mathbb{R}^{N\times m}$ contains the features, $Y \in \mathbb{R}^{N\times 1}$ is the target variable, and $\hat{\beta} \in \mathbb{R}^{m\times 1}$ are the regression coefficients.

The closed-form solution is obtained by minimizing the mean squared error:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Using the full dataset ($N = 1030$) with 8 input features plus a bias term (9 parameters total), we obtained the following coefficients:

$$\hat{\beta} = \begin{bmatrix} -23.33 & 0.12 & 0.10 & 0.09 & -0.15 & 0.29 & 0.018 & 0.020 & 0.11 \end{bmatrix}^T.$$

The in-sample mean squared error (MSE) was approximately

$$\text{MSE}_{\text{train}} \approx 107.20.$$

Q1.2 Feature Subset Comparison

We randomly split the dataset into 75% training (772 samples) and 25% validation (258 samples). We trained models with 9, 8, and 7 parameters (bias plus features) by progressively removing features. The validation MSEs are reported below:

Model	Validation MSE
9 features (all)	103.97
8 features (drop flyash)	104.38
7 features (drop flyash + slag)	114.90

Observations

- The model with all 9 parameters achieved the lowest validation error.
- Removing one feature (`flyash`) only slightly degraded performance.
- Removing two features (`flyash` and `slag`) caused a significant increase in error.
- This suggests that all features provide useful information and, in this case, reducing the feature set led to worse generalization. Unlike some cases where more features can cause overfitting, here the full feature set gave the best results.

Problem 2: Multinomial Logistic Regression with Pre-trained Feature Extractor

Model Setup

We used the provided pre-trained CNN feature extractor to obtain latent representations $h \in \mathbb{R}^{256}$ from each MNIST image $x \in \mathbb{R}^{784}$. On top of these frozen features, we trained a multinomial logistic regression (softmax) classifier.

The prediction rule is

$$\hat{y} = \arg\max_c \sigma(W^T h + b),$$

where $W \in \mathbb{R}^{256\times 10}$, $b \in \mathbb{R}^{1\times 10}$, and $\sigma(\cdot)$ denotes the softmax function.

Gradient Derivation

For a mini-batch of size B, the cross-entropy loss is

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{10} y_{ic}\,\log \hat{p}_{ic}, \qquad \hat{p} = \sigma(W^T h + b).$$

Gradients are

$$\frac{\partial \mathcal{L}}{\partial W} = \frac{1}{B} H^\top (\hat{P} - Y), \qquad \frac{\partial \mathcal{L}}{\partial b} = \frac{1}{B}\mathbf{1}^\top (\hat{P} - Y),$$

where $H \in \mathbb{R}^{B\times 256}$ is the feature matrix, $\hat{P}$ is the predicted probability matrix, and Y is the one-hot encoded label matrix.

Training Procedure

We implemented stochastic gradient descent (SGD) using only matrix operations (no autograd or built-in optimizers). Weights were updated as

$$W \leftarrow W - \eta\,\frac{\partial \mathcal{L}}{\partial W}, \qquad b \leftarrow b - \eta\,\frac{\partial \mathcal{L}}{\partial b}.$$

Results

We trained the model for 30 epochs with learning rate $\eta = 0.1$ and batch size 512. The training and validation losses consistently decreased and converged smoothly. Validation accuracy reached **99.97**%, and test accuracy reached **99.07**%.

- Training loss decreased from 0.1469 at epoch 1 to 0.0050 at epoch 30.
- Validation loss decreased from 0.0467 at epoch 1 to 0.0054 at epoch 30.
- The model achieved near-perfect classification performance using the frozen extractor.

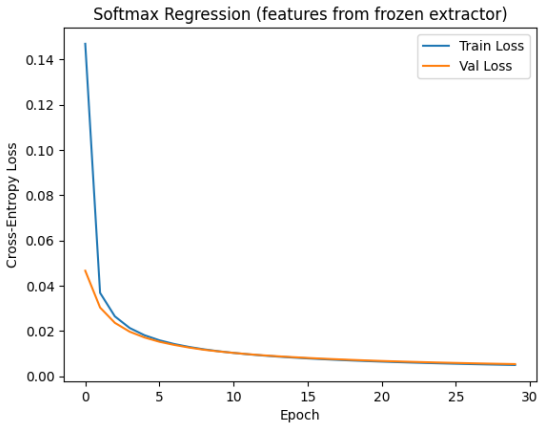


Figure 1: Training and validation losses.

Conclusion

Logistic regression trained on top of pre-trained features can achieve very high accuracy on MNIST. The feature extractor effectively reduces dimensionality and provides separable latent representations, allowing the linear classifier to converge quickly and generalize well.

Problem 3: Transformation from a Uniform Distribution

Let $U_1, U_2 \overset{\text{i.i.d.}}{\sim} \text{Unif}(0,1)$ and define

$$\Theta = 2\pi U_1, \qquad R = \sqrt{-2\ln U_2}, \qquad Z_1 = R\cos\Theta, \quad Z_2 = R\sin\Theta.$$

Step 1: Distributions of R and Θ. Since $U_1 \sim \text{Unif}(0,1)$, we have $\Theta \sim \text{Unif}(0, 2\pi)$. For $R = \sqrt{-2\ln U_2}$,

$$\mathbb{P}(R \le r) = \mathbb{P}(U_2 \ge e^{-r^2/2}) = 1 - e^{-r^2/2}, \quad r \ge 0,$$

so $f_R(r) = r\, e^{-r^2/2}\, \mathbf{1}_{\{r\ge 0\}}$ (Rayleigh). Because U_1 and U_2 are independent, R and Θ are independent and

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi}\, re^{-r^2/2}\, \mathbf{1}_{\{r\ge 0,\ \theta\in[0,2\pi)\}}.$$

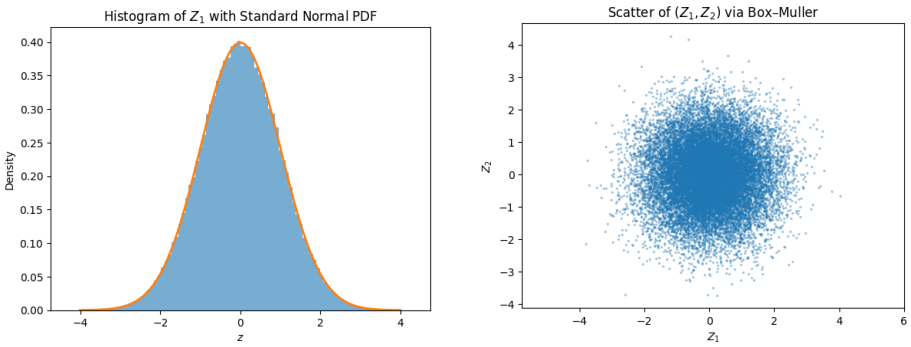
Step 2: Transform to (Z_1, Z_2). Using the polar map $(z_1, z_2) \mapsto (r, \theta)$ with $r = \sqrt{z_1^2 + z_2^2}$ and $\theta = \text{atan2}(z_2, z_1)$, the Jacobian satisfies $|\partial(r,\theta)/\partial(z_1,z_2)| = \frac{1}{r}$ for $r > 0$. Thus the joint pdf of (Z_1, Z_2) is

$$f_{Z_1,Z_2}(z_1,z_2) = f_{R,\Theta}(r,\theta)\cdot\frac{1}{r} = \frac{1}{2\pi}\exp\left(-\frac{z_1^2+z_2^2}{2}\right), \quad (z_1,z_2)\in\mathbb{R}^2.$$

Step 3: Standard normal marginals and independence. Since

$$f_{Z_1,Z_2}(z_1,z_2) = \left(\tfrac{1}{\sqrt{2\pi}}e^{-z_1^2/2}\right)\left(\tfrac{1}{\sqrt{2\pi}}e^{-z_2^2/2}\right) = \phi(z_1)\,\phi(z_2),$$

each marginal equals $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$; hence $Z_1, Z_2 \sim \mathcal{N}(0,1)$, and the factorization shows Z_1 and Z_2 are independent.



(a) Histogram of Z_1 overlaid with $\mathcal{N}(0,1)$ PDF.

(b) Scatter of (Z_1, Z_2) showing circular symmetry (independence).

Figure 2: Empirical validation of the Box–Muller transform.

Empirical Verification of Box–Muller

We generated $n = 200{,}000$ i.i.d. pairs $(U_1, U_2) \sim \mathrm{Unif}(0,1)^2$ and applied $\Theta = 2\pi U_1$, $R = \sqrt{-2\ln U_2}$, $Z_1 = R\cos\Theta$, $Z_2 = R\sin\Theta$. The resulting samples satisfy

$$\hat{\mu}_{Z_1} = -7\times 10^{-4}, \quad \widehat{\mathrm{Var}}(Z_1) = 1.0017,$$
$$\hat{\mu}_{Z_2} = -6\times 10^{-4}, \quad \widehat{\mathrm{Var}}(Z_2) = 1.0042,$$
$$\widehat{\mathrm{corr}}(Z_1,Z_2) = 0.0004\ .$$

The histogram of Z_1 closely matches the standard normal density and the scatter of (Z_1, Z_2) is approximately circular, confirming that $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Together with the analytic Jacobian derivation above, this verifies both the standard normal marginals and independence.

Conclusion. The Box–Muller transform produces two independent standard normal variables from two independent uniforms.