8 Measuring Performance

Problem 8.1 Will the multiclass cross-entropy training loss in figure 8.2 ever reach zero? Explain your reasoning.

Given that each training example is different, the model could perform perfect inferencing of the training set if the model has enough capacity to describe the training data.

Problem 8.2 What values should we choose for the three weights and biases in the first layer of the model in figure 8.4a so that the hidden unit’s responses are as depicted in figures 8.4b-d?

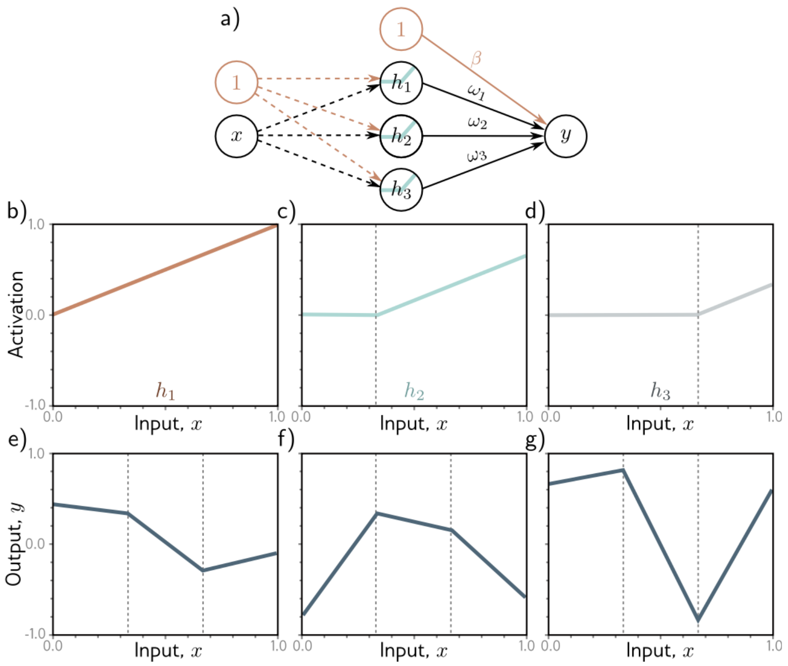
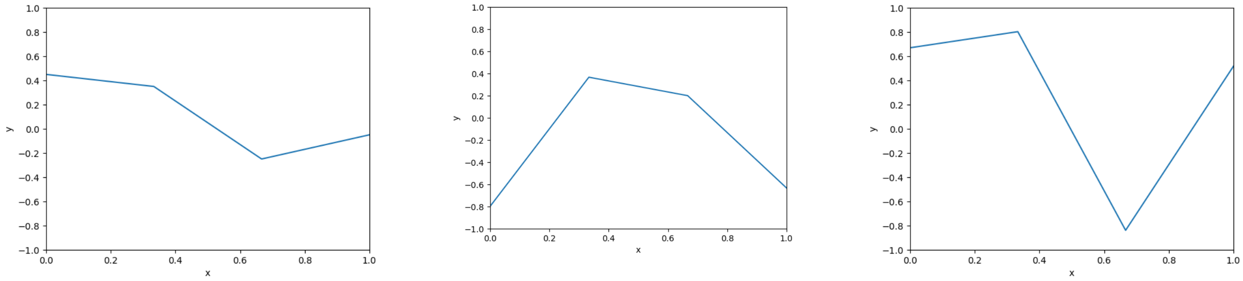


Figure 8.4 Simplified neural network with three hidden units. a) The weights and biases between the input and hidden layer are fixed (dashed arrows). b–d) They are chosen so that the hidden unit activations have slope one, and their joints are equally spaced across the interval, with joints at $x = 0$, $x = 1/3$, and $x = 2/3$, respectively. Modifying the remaining parameters $\boldsymbol{\phi} = \{\beta, \omega_1, \omega_2, \omega_3\}$ can create any piecewise linear function over $x \in [0, 1]$ with joints at $1/3$ and $2/3$. e–g) Three example functions with different values of the parameters $\boldsymbol{\phi}$.

```
# %% [markdown]
# Given a linear regression with three hidden units at fixed jo

# %%
import numpy as np
import matplotlib.pyplot as plt

# %%
def ReLU(h):
  return h.clip(0.0)

# Three fixed weights and biases for first layer
def h1(x):
  return ReLU(x) # joint at x=0

def h2(x):
  return ReLU(x-1/3) # joint at x=1/3

def h3(x):
  return ReLU(x-2/3) # joint at x=2/3

def model(x, w1, w2, w3, b):
  return b + w1*h1(x) + w2*h2(x) + w3*h3(x)

# %%
def plot(w1, w2, w3, b):
  x = np.linspace(0, 1, 100)
  y = model(x, w1, w2, w3, b)
  plt.plot(x, y)
  plt.xlabel('x')
  plt.ylabel('y')
  plt.xlim(0, 1)
  plt.ylim(-1, 1)
  plt.yticks(np.arange(-1, 1.2, 0.2))
  plt.show()
```

```
# Now search for weights and biases that make the model produces
plot(-0.3, -1.5, 2.4, 0.45)
plot(3.5, -4, -2, -0.8)
plot(0.4, -5.33, 9, 0.67)
```



Problem 8.3 Given a training dataset consisting of I input/output pairs $\{x_i, y_i\}$, show how the parameters $\{\beta, \omega_1, \omega_2, \omega_3\}$ for the model in figure 8.4a using the least squares loss function can be found in closed form.

Note that h_1, h_2, h_3 for each training example can be found with fixed weights and biases:

$$
\begin{aligned}
h_{1,i} &= \mathrm{ReLU}(x_i) \\
h_{2,i} &= \mathrm{ReLU}(x_i - 1/3) \\
h_{3,i} &= \mathrm{ReLU}(x_i - 2/3)
\end{aligned}
$$

Then the loss function for the i-th training example is computed as

$$
l_i = (y_i - (\beta + \omega_1 h_{1,i} + \omega_2 h_{2,i} + \omega_3 h_{3,i}))^2
$$

Problem 8.4 Consider the curve in figure 8.10b at the point where we train a model with a hidden layer of size 200, which would have 50,410 parameters. What do you predict will happen to the training and test performance if we increase the number of training examples from 10,000 to 50,410?

As the model consists of more hidden units, the model has enough parameters to memorize the entire dataset. With 50,410 parameters, the model has enough capacity to represent the training set of 50,410 examples. Thus, the training loss will remain zero.

However, the performance in the test set will worsen although more training examples are provided. Now the model performance lies on the *critical regime*, where the test error increases due to overfitting.

Problem 8.5 Consider the case where the model capacity exceeds the number of training data points, and the model is flexible enough to reduce the training loss to zero. What are the implications of this for fitting a heteroscedastic model? Propose a method to resolve any problems you identify.

We can predict possible problems in perspective of bias-variance tradeoff. For heteroscedastic models, the inherent variance in the training data is different accross output dimensions. Thus, it is possible for the model to overfit to only some of the outputs, complicating the hyperparameterssearch to find the adequate capacity for the model. To address this, we could apply normalization to the output layer to transform the problem into fitting a homoscedastic model.

Problem 8.6 Show that two random points drawn from a 1000-dimensional standard Gaussian distribution are orthogonal relative to the origin with high probability.

Let's generalize this problem into evaluating the orthogonality of two D-dimensional vectors $\mathbf{u}, \mathbf{v}$ randomly drawn from standard Gaussian distribution.

$$\mathbf{u} \cdot \mathbf{v} = \sum_{D} u_i v_i$$

Since $u_i, v_i \sim \mathcal{N}(0, 1^2)$ and are independent, $u_i v_i \sim \mathcal{N}(0, 1^2)$ and thus $\mathbf{u} \cdot \mathbf{v} \sim \mathcal{N}(0, D^2)$. Therefore, we can say that

$$\frac{||\mathbf{u}\cdot\mathbf{v}||^2}{||\mathbf{u}||^2||\mathbf{v}||^2} \approx \frac{D^2}{D^2D^2} = \frac{1}{D^2}$$

which indicates that as D increases, the vectors $\mathbf{u}$ and $\mathbf{v}$ becomes *almost* orthogonal.

Problem 8.7 The volume of a hypersphere with radius r in D dimensions is:

$$\mathrm{Vol}[r] = \frac{r^D\pi^{D/2}}{\Gamma[D/2+1]}$$

where $\Gamma[\bullet]$ is the Gamma function. Show using Stirling's formula that the volume of a hypersphere of diameter one (radius $r = 0.5$) becomes zero as the dimension increases.

The gamma function according to the Stirling's formula can be approximated to

$$\Gamma[z+1] \approx \sqrt{2\pi z}(\frac{z}{e})^z$$

Thus,

$$\begin{aligned}\mathrm{Vol}[r] &\approx \frac{r^D\pi^{D/2}}{\sqrt{\pi D}(\frac{D/2}{e})^{D/2}}\\ &= \frac{1}{\sqrt{\pi D}}r^D(\frac{2\pi e}{D})^{D/2}\end{aligned}$$

which approaches zero as D increases, for any $r < 1$.

Problem 8.8 Consider a hypersphere of radius $r = 1$. Find an expression for the proportion of the total volume that lies in the outermost 1% of the distance from the center (i.e. in the outermost shell of thickness 0.01). Show that this becomes one as the dimension increases.

Using the formula above,

$$\text{Vol}[r=1] = \frac{\pi^{D/2}}{\Gamma[D/2+1]}$$

and

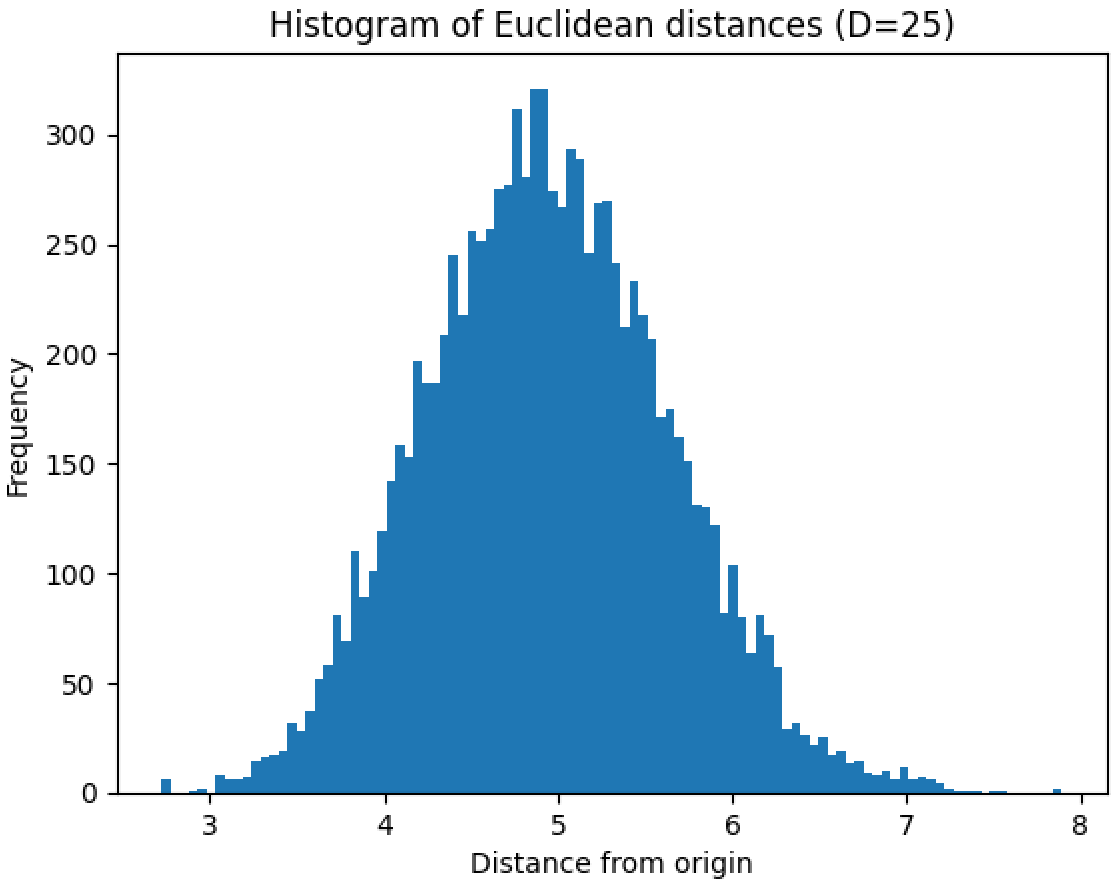
$$\text{Vol}[r=0.99] = \frac{0.99^D \pi^{D/2}}{\Gamma[D/2+1]}$$

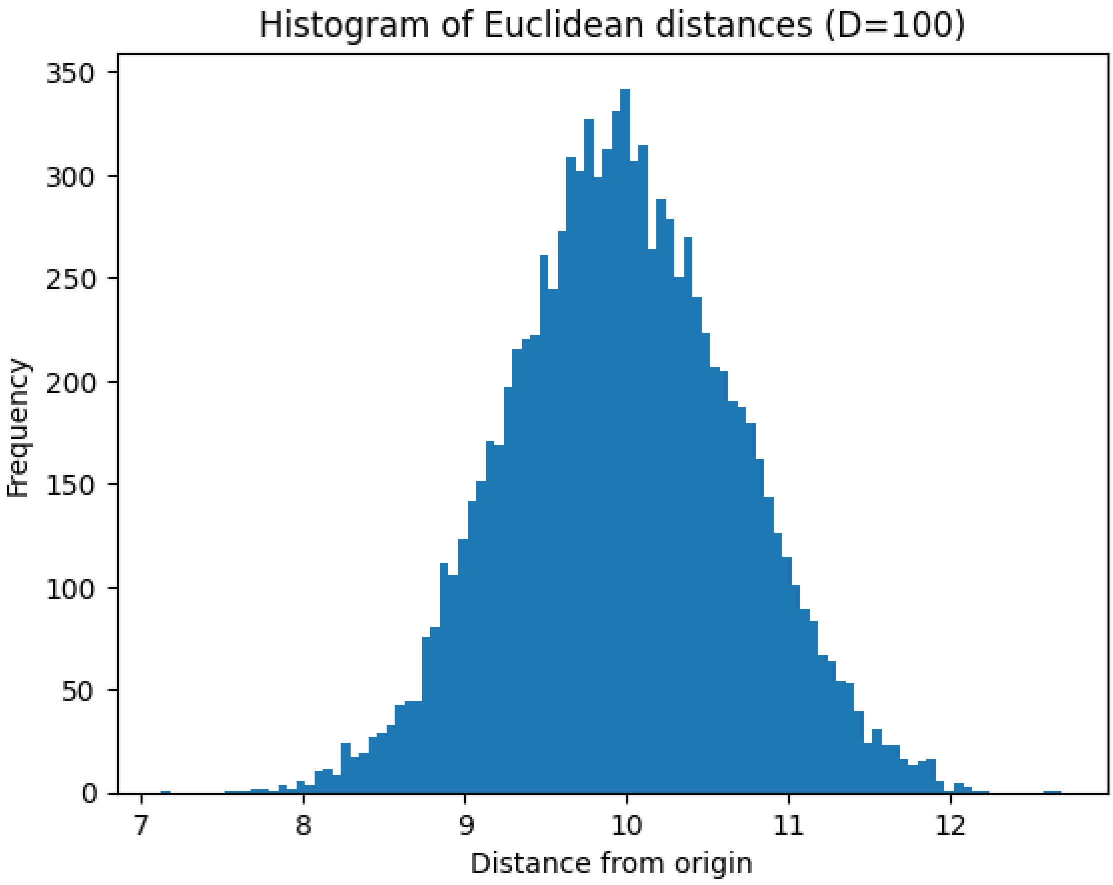
Thus the proportion of the volume of shell is:

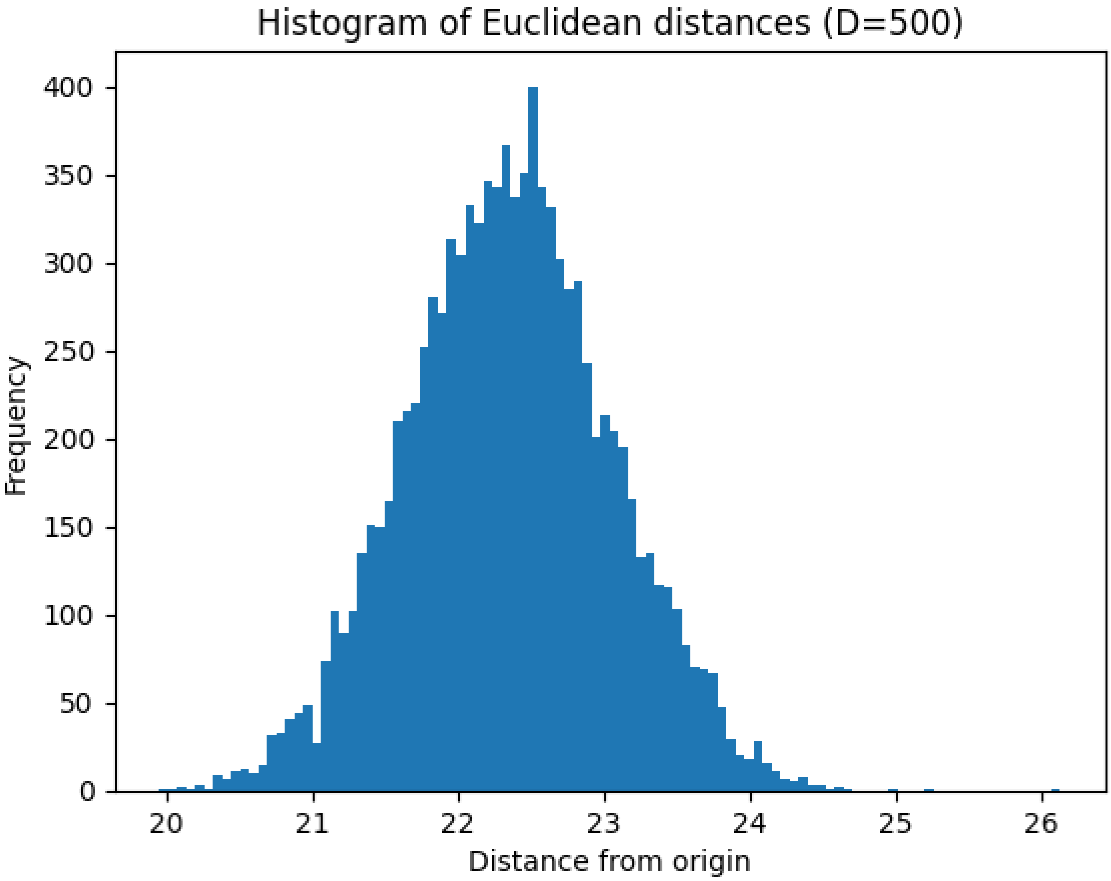
$$\frac{\text{Vol}[r=1]-\text{Vol}[r=0.99]}{\text{Vol}[r=1]} = 1-0.99^D$$

which approaches 1 as $D \rightarrow \infty$.

Problem 8.9 Figure 8.13c shows the distribution of distances of samples of a standard normal distribution as the dimension increases. Empirically verify this finding by sampling from the standard normal distributions in 25, 100, and 500 dimensions and plotting a histogram of the distances from the center. What closed-form probability distribution describes these distances?


Histogram of Euclidean distances (D=25)
Frequency
300
250
200
150
100
50
0
3
4
5
6
7
8
Distance from origin


Histogram of Euclidean distances (D=100)
Frequency
350
300
250
200
150
100
50
0
7
8
9
10
11
12
Distance from origin



We can see that the distribution of the distance from the origin follows the normal distribution $\mathcal{N}(\sqrt{D}, 1)$.