Question 1

Model Architecture:

The model would consist of a generator and discriminator:

The generator would try and generate "replicate" the bold 1s in the dataset.

It's input would be an image of regular 1, as well as a random noise vector (could be sampled from normal distribution),

And it would be trained to generate a "bold" versions of the 1.

The discriminator would try and distinguish between the real "bold" 1s and the generated "bold" 1s.

It's input would be the generated image.

And it would be trained to distinguish between real "bold" 1s and generated images, by outputting a Boolean flag True if the image is real and False if its not.

The specific architecture and the parameters of the model can be optimized when building such model.

Training procedure:

The loss function for the generator could be the **binary cross-entropy loss** between the generated images and the real "bold" 1s.

And the loss for the discriminator is the **binary cross-entropy loss** between the predicted labels for the generated images and the real "bold" 1s.

It would be trained on a real "bold" 1 with the label True (1) and then trained on a generated image with label False (0).

The training scheme would be updating the generator for multiple iterations for each update of the discriminator.

Generation method:

The generation method to fulfill the task would be feeding and image of regular 1 and a random noise vector into the generator and use the output as a "bold" version of the 1.

Assumptions:

One assumption would be that the underlying patterns and features of the "bold" 1s images can be learned by the generator.

Other assumption would be that these patterns and features can be used to generate a new "bold" 1s that are similar to the ones in the training set.

Limitations:

For the training procedure to be successful the model (Generator and Discriminator) would need sufficient amount of “bold” 1s to learn it’s features.

And the model’s ability to perform well on images that’s not in the training set, and generate “bold” 1s that are different from the training set.

Additionally the model may not generalize well to images that’s not in the MNIST dataset and may not perform well when the training set is imbalanced.

Question 2

a

From the definition of the Rademacher Complexity we get:

$$R(\hat{F} \circ S) = \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m}[sup_{\hat{f} \in \hat{F}} \sum_{i=1}^{m} \sigma_i \hat{f}(z_i)]$$

$$= \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m}[sup_{f \in F} \sum_{i=1}^{m} \sigma_i c \cdot f(z_i)]$$

case 1 $c \geq 0$:

$$^1 = \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m}[c \cdot sup_{f \in F} \sum_{i=1}^{m} \sigma_i f(z_i)]$$

$$^2 = c \cdot \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m}[sup_{f \in F} \sum_{i=1}^{m} \sigma_i f(z_i)]$$

$$= c \cdot R(F \circ S) = |c| \cdot R(F \circ S)$$

case 2 $c < 0$:

$$^3 = \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m}[|c| \cdot sup_{f \in F} \sum_{i=1}^{m} -\sigma_i f(z_i)]$$

$$^4 = |c| \cdot \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m}[sup_{f \in F} \sum_{i=1}^{m} -\sigma_i f(z_i)]$$

$$= |c| \cdot R(F \circ S)$$

[1] Linearity of sup for positive integers.

[2] Linearity of Expectation.

[3] Linearity of sup for positive numbers. $|c|$

[4] Since $\sigma \in \{\pm 1\}$ multiplying the σ by -1 doesn't change the distribution and so doesn't change the value of the Expectation

b

Denote $\hat{F_1}$ to be the $\hat{F}$ defined in part a

From the definition of the Rademacher Complexity we get:

$$R(\hat{F} \circ S) = \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m} [sup_{\hat{f} \in \hat{F}} \sum_{i=1}^{m} \sigma_i \hat{f}(z_i)]$$

$$^1 = \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m} [sup_{f \in F} \sum_{i=1}^{m} \sigma_i (c \cdot f(z_i) + b)]$$

$$^2 = \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m} [sup_{f \in F} \sum_{i=1}^{m} \sigma_i c \cdot f(z_i) + \sum_{i=1}^{m} \sigma_i b]$$

$$^3 = \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m} [sup_{f \in F} \sum_{i=1}^{m} \sigma_i c \cdot f(z_i)] + \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m} [\sum_{i=1}^{m} \sigma_i b]$$

$$^4 = R(\hat{F_1} \circ S) + b \sum_{i=1}^{m} \frac{1}{m} E_{\sigma \sim \{\pm 1\}^m} [\sigma_i]$$

$$^5 = R(\hat{F_1} \circ S)$$

$$^6 = |c| \cdot R(F \circ S)$$

---

[1] substitute $\hat{f}$ by c$\cdot f + b$

[2] $\sum_{i=1}^{m} \sigma_i b$ isn't dependant on f so can be taken out of the sup

[3] Linearity of Expectation

[4] Definition of $R(\hat{F_1} \circ S)$ and [3]

[5] $\mathrm{E}[\sigma_i] = 0, \forall i$

[6] from a