SET A

Q1

Variance measures the extent to which the learnt model is sensitive to a particular training dataset of a finite size. High variance indicates overfitting.

$$variance = \int \mathbb{E}_{\mathcal{D}}[\{y(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]\}^2]p(x)\,dx$$

Q2

Key difference:
Huber loss:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

The error is the squared loss if the residual ($|y - f(x)|$) is less than a threshold δ and is the absolute loss otherwise. Hence, the key difference is that for large residuals, the squared loss imposes a squared penalty (quadratic) and the Huber loss imposes only the absolute penalty (linear).
Robust to outliers:
When the residuals are greater than 1, the squared penalty is more than the absolute penalty. Hence, Huber loss is more robust to outliers since it imposes a lesser penalty on the outliers which have large residuals (when compared to squared loss). Thus the model is less "influenced" by the outliers.

Q3

Two reasons why we are able to evade the curse-of-dimensionality for this high-dimensional dataset:

1. The data exists on a lower dimensional manifold of this high dimensional space. We saw in class that MNIST can be represented in lower dimensions without significant loss, this shows that the data exists in a lower dimensional manifold of the higher dimensional space. For example, the points on a line in a 3 dimensional space are effectively on a 1 dimensional subspace. (this reason is sufficient to receive full credit).

2. The input variables (the images) exhibits smoothness properties - small changes in the input variables will not result in the label (the target variable) changing. Hence, if we get a new input variable which is close to a particular input variable which the model has trained on, it can predict the corresponding label.

(No credit is awarded for explaining PCA.)

Q4

Refer chapter 19 in KJ.

SET B

Q1

Variance measures the extent to which the learnt model is sensitive to a particular training dataset of a finite size. High variance indicates overfitting.

$$variance = \int \mathbb{E}_{\mathcal{D}}[\{y(x;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]\}^2]p(x)\,dx$$

Q2

Huber loss:

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \leq \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

The error is the squared loss if the residual ($|y - f(x)|$) is less than a threshold δ and is the absolute loss otherwise. Hence, the error is down-weighted for large residuals - the squared loss imposes a squared penalty (quadratic) and the Huber loss imposes only the absolute penalty (linear).
When the residuals are greater than 1, the squared penalty is more than the absolute penalty. Hence, Huber loss is more robust to outliers since it imposes a lesser penalty on the outliers which have large residuals (when compared to squared loss). Thus the model is less “influenced” by the outliers.

Q3

Two reasons why we are able to evade the curse-of-dimensionality for this high-dimensional dataset:

1. The data exists on a lower dimensional manifold of this high dimensional space. For example, the points on a line in a 3 dimensional space are effectively on a 1 dimensional subspace. (this reason is sufficient to receive full credit).

2. The input variables (the images) exhibits smoothness properties - small changes in the input variables will not result in the label (the target variable) changing. Hence, if we get a new input variable which is close to a particular input variable which the model has trained on, it can predict the corresponding label.

(No credit is awarded for explaining PCA.)

Q4

Refer chapter 19 in KJ.