Our training data is X, a $n \times p$ matrix. Each example therefore has shape $p \times 1$[1] and is labeled with either a $+1$ or a -1. The two classes have means $\mathbf{m}_1$ and $\mathbf{m}_2$ (so they also have shape $p \times 1$). For the MNIST data set, $n = 60k$, the number of training examples, and $p = 28 \times 28 = 784$, the number of pixels in each image. We assume there are n_1 examples with label 1 and n_2 examples with label -1

We project all data points onto a linear space of dimension 1 (a line) in the direction of the unit vector $\mathbf{u}$. We focus on two metrics that together determine how good this linear space is for classification.

Scatter between classes As we saw in class, the larger the distance between the class means along the unit vector $\mathbf{u}$, the more we like it. Furthermore, the squared distance between the class means $\mathbf{m}_1$ and $\mathbf{m}_2$ along $\mathbf{u}$ is

$$\left(\mathbf{u}^T(\mathbf{m}_1 - \mathbf{m}_2)\right)^2 = \mathbf{u}^T S_b \mathbf{u}.$$

Therefore, we want to maximize $\mathbf{u}^T S_b \mathbf{u}$. But this by itself is not enough, since the direction maximizing the projections of the means may also have a large spread/variance within each class (hence large overlap between the two classes).

Scatter within classes So we came up with the idea of controlling the scattering within each class. To quantify this, we defined the intra-class scatter of any class (say, class 1) to be

$$\mathbf{u}^T \left(\sum_{\mathbf{x}:\text{label of } \mathbf{x}=1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^T \right) \mathbf{u}.$$

The total scatter for all classes is the sum of the scatter over each class, which we showed to be $\mathbf{u}^T S_w \mathbf{u}$, where

$$S_w = \sum_{\mathbf{x}:\text{label of } \mathbf{x}=1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^T + \sum_{\mathbf{x}':\text{label of } \mathbf{x}'=-1} (\mathbf{x}' - \mathbf{m}_2)(\mathbf{x}' - \mathbf{m}_2)^T.$$

Now observe that S_w is also a $p \times p$ matrix. So long as there are p linearly independent examples among the n examples, we will have from our previous section on rank that the rank of S_w is p. Therefore, in what follows, we will therefore consider S_w to be a full rank matrix. The eigenvectors of S_w cannot be read out from the expression above like we did for S_b, but

[1] We name only column vectors, so if $\mathbf{x}_1$ is the first example (shape $p \times 1$), the first row of X is $\mathbf{x}_1^T$ with shape $1 \times p$

note that the direction that maximizes $\mathbf{u}^T S_w \mathbf{u}$ remains the eigenspace of the largest eigenvalue, and the direction minimizing $\mathbf{u}^T S_w \mathbf{u}$ is the eigenspace of the smallest eigenvalue.

Recall from last class: S_b has rank 1, and its eigenvalues are $|\mathbf{m}_1 - \mathbf{m}_2|^2$ and $p-1$ zeros. The eigenspace corresponding to the eigenvalue $|\mathbf{m}_1 - \mathbf{m}_2|^2$ is simply the linear space in the direction of $\mathbf{m}_1 - \mathbf{m}_2$.

We will simplify S_w a little more. Note that

$$\begin{aligned}
S_w &= \sum_{\mathbf{x}\text{:label of } \mathbf{x}=1} (\mathbf{x}-\mathbf{m}_1)(\mathbf{x}-\mathbf{m}_1)^T + \sum_{\mathbf{x}'\text{:label of } \mathbf{x}'=-1} (\mathbf{x}'-\mathbf{m}_2)(\mathbf{x}'-\mathbf{m}_2)^T \\
&= \sum_{\mathbf{x}\text{:label of } \mathbf{x}=1} (\mathbf{x}\mathbf{x}^T - \mathbf{x}\mathbf{m}_1^T - \mathbf{m}_1\mathbf{x}^T + \mathbf{m}_1\mathbf{m}_1^T) \\
&\qquad + \sum_{\mathbf{x}'\text{:label of } \mathbf{x}'=-1} (\mathbf{x}'\mathbf{x}'^T - \mathbf{x}'\mathbf{m}_2^T - \mathbf{m}_2\mathbf{x}'^T + \mathbf{m}_2\mathbf{m}_2^T) \\
&\overset{(a)}{=} \left(\sum_{\mathbf{x}\text{:label of } \mathbf{x}=1} (\mathbf{x}\mathbf{x}^T) \right) - n_1\mathbf{m}_1\mathbf{m}_1^T - n_1\mathbf{m}_1\mathbf{m}_1^T + n_1\mathbf{m}_1\mathbf{m}_1^T \\
&\qquad + \left(\sum_{\mathbf{x}'\text{:label of } \mathbf{x}'=-1} (\mathbf{x}'\mathbf{x}'^T) \right) - n_2\mathbf{m}_2\mathbf{m}_2^T - n_2\mathbf{m}_2\mathbf{m}_2'^T + n_2\mathbf{m}_2\mathbf{m}_2^T \\
&= \left(\sum_{\mathbf{x}\text{:label of } \mathbf{x}=1} (\mathbf{x}\mathbf{x}^T) + \sum_{\mathbf{x}'\text{:label of } \mathbf{x}'=-1} (\mathbf{x}'\mathbf{x}'^T) \right) - n_1\mathbf{m}_1\mathbf{m}_1^T - n_2\mathbf{m}_2\mathbf{m}_2^T,
\end{aligned}$$

where (a) follows since

$$\mathbf{m}_1 = \left(\sum_{\mathbf{x} \text{ with label } 1} \mathbf{x} \right)/n_1 \text{ and } \mathbf{m}_2 = \left(\sum_{\mathbf{x}' \text{ with label } -1} \mathbf{x}' \right)/n_2$$

Also from the usual outer-product way of multiplying matrices, we have

$$\left(\sum_{\mathbf{x}\text{:label of } \mathbf{x}=1} (\mathbf{x}\mathbf{x}^T) + \sum_{\mathbf{x}'\text{:label of } \mathbf{x}'=-1} (\mathbf{x}'\mathbf{x}'^T) \right) = X^T X,$$

implying that

$$S_w = X^T X - n_1\mathbf{m}_1\mathbf{m}_1^T - n_2\mathbf{m}_2\mathbf{m}_2^T.$$

Exercise Suppose X is centered, namely the rows of X add up to $\mathbf{0}$. In other words, you have $n_1\mathbf{m}_1 + n_2\mathbf{m}_2 = \mathbf{0}$. Show then, that

$$\frac{n_1 n_2}{n_1+n_2}(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T = n_1\mathbf{m}_1\mathbf{m}_1^T + n_2\mathbf{m}_2\mathbf{m}_2^T.$$

Hint: note that $n_1\mathbf{m}_1 = -n_2\mathbf{m}_2$ implies that

$$\mathbf{m}_1 - \mathbf{m}_2 = \frac{n_1 + n_2}{n_2}\mathbf{m}_1 = -\frac{n_1 + n_2}{n_1}\mathbf{m}_2.$$

Therefore, we have $S_w + \frac{n_1 n_2}{n_1+n_2} S_b = X^T X$.

Posing the Fisher Discriminant problem The optimal direction is the solution of the equation

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}.$$

Recall: the greater the separation between the means in the direction $\mathbf{w}$, the higher the numerator. The *smaller* the variance/scatter within classes, the smaller the denominator. Ideally, we want directions that increase the numerator, but also ones in which the denominator reduces—that is why we maximize the ratio of the two.

The expression on the right above is impervious to scaling of $\mathbf{w}$. Normally, we scale $\mathbf{w}$ to have length 1 in cases like this, but here we do something different. We scale different directions by different amounts, and in particular, we scale in the direction of $\mathbf{w}$ just enough that the denominator becomes 1.

Doing so, we can rewrite the objective as finding

$$\begin{aligned} \mathbf{w}^* &= \arg\max_{\mathbf{w}} \mathbf{w}^T S_b \mathbf{w} \\ \text{subject to } & \mathbf{w}^T S_w \mathbf{w} = 1. \end{aligned}$$

This is a constrained optimization problem, which we solve using the idea of a Lagrangian. This is actually the third constrained problem we have encountered thus far in this class alone (the prior two times, we just sneakily went ahead without saying much).