Homework 10 - Solution

1. (a) By definition,

$$(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})_{ij} = \sum_{k=1}^{n} \tilde{\mathbf{X}}_{ki} \tilde{\mathbf{X}}_{kj} = \sum_{k=1}^{n} \Phi(\mathbf{x_k})_i \Phi(\mathbf{x_k})_j.$$

Recall that the vector $\Phi(\mathbf{x_k})$ contains zero in all of its entries but one. Therefore, if $i \neq j$, $\Phi(\mathbf{x_k})_i \Phi(\mathbf{x_k})_j = 0$, so that $(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})_{ij} = 0$. This means that non-diagonal entries of $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ are equal to zero. For $i = j$, we obtain

$$(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})_{ii} = \sum_{k=1}^{n} \Phi(\mathbf{x_k})_i^2.$$

Having $\Phi(\mathbf{x_k})_i = 1$ means that $\mathbf{x_k} \in A_i$. Therefore, this sum is equal to the number of observations $\mathbf{x_k}$ in A_i, that is $\mathbf{n_i}$. This is enough to conclude.

(b) If we assume that all the $\mathbf{n_j}$s are positive, then $(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}$ is the diagonal matrix with entries $\mathbf{n_j}^{-1}$. Let us compute $\tilde{\mathbf{X}}^\top \mathbf{Y}$:

$$\tilde{\mathbf{X}}^\top \mathbf{Y} = \sum_{i=1}^{n} \mathbf{y_i} \Phi(\mathbf{x_i}) = \begin{pmatrix} \sum_{i=1}^{n} \mathbf{y_i} \mathbf{1}\{\mathbf{x_i} \in A_1\} \\ \vdots \\ \sum_{i=1}^{n} \mathbf{y_i} \mathbf{1}\{\mathbf{x_i} \in A_J\} \end{pmatrix}.$$

Therefore, $\hat{a} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y}$ is equal to

$$\hat{a} = \begin{pmatrix} \frac{1}{\mathbf{n_1}} \sum_{i=1}^{n} \mathbf{y_i} \mathbf{1}\{\mathbf{x_i} \in A_1\} \\ \vdots \\ \frac{1}{\mathbf{n_J}} \sum_{i=1}^{n} \mathbf{y_i} \mathbf{1}\{\mathbf{x_i} \in A_J\} \end{pmatrix}.$$

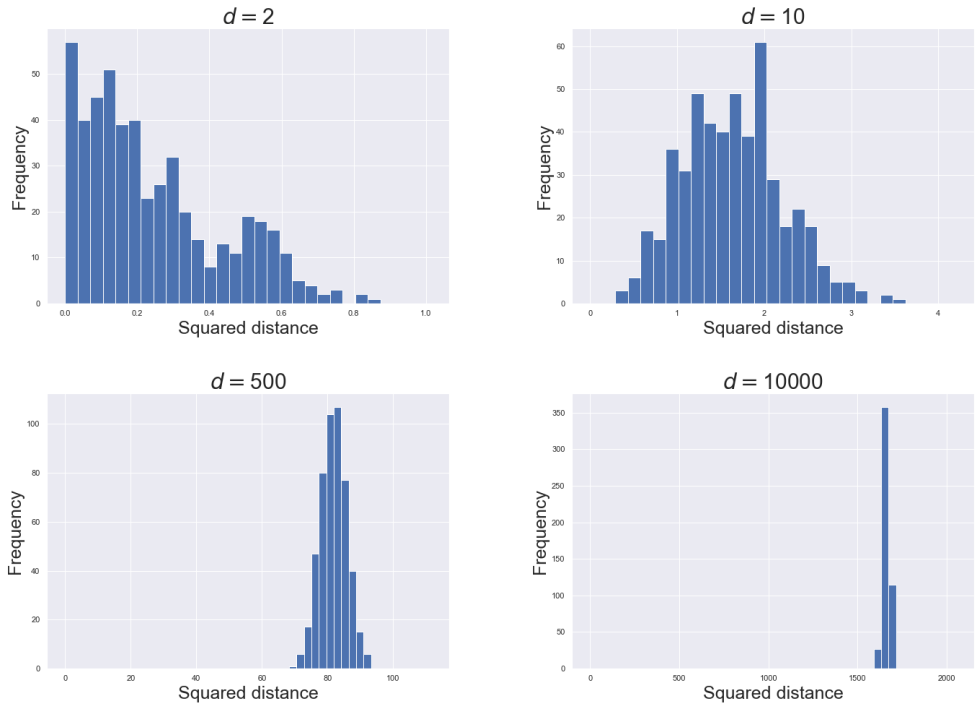

Figure 1: Distribution of the squared distances in different dimensions.

Eventually, $\hat{f}_{\mathrm{LS}}(x)$ is equal to $\langle \hat{a}, \Phi(x) \rangle$. Assume that $x \in A_j$. In this case, the vector $\Phi(x)$ contains 0 everywhere, and a 1 at jth entry. Therefore, $\langle \hat{a}, \Phi(x) \rangle = \frac{1}{\mathbf{n_j}} \sum_{i=1}^{n} \mathbf{y_i} \mathbf{1}\{\mathbf{x_i} \in A_j\}$. This is equal to $\hat{f}_{\mathcal{A}}(x)$.

2. We display in Figure 1 the histograms of the squared distances to the point $\mathbf{x}$ for different dimensions d. There are several observations to be made. First, the average distance increases with d: from 0.24 for $d = 2$ to 1660 for $d = 10000$. The standard deviation also increases but at a much slower rate: from 0.19 for $d = 1$ to 17 for $d = 10000$. Therefore, in small dimension, the standard deviation is roughly equal to the expectation. But in high dimension, the distribution is very concentrated around its expectation, as the standard deviation is much smaller than the expectation.

3. As the distribution is very concentrated, it means that with high probability, all the distances $\|\mathbf{x} - \mathbf{x_i}\|$ are roughly equal. If all the distances are the same, we cannot hope that knowing what the nearest neighbor of $\mathbf{x}$ is gives us any meaningful information about $\mathbf{x}$. This explains why nearest-neighbor methods fail in high dimension.