Problem Solutions

e-Chapter 8

Pierre Paquay

Problem 8.1

The two separation constraints are

$$(w^T x_+ + b) \geq 1 \text{ and } -(w^T x_- + b) \geq 1;$$

by adding these two constraints, we get that

$$w^T(x_+ - x_-) \geq 2.$$

Then, the Cauchy-Schwarz inequality gives us the following inequalities

$$2 \leq w^T(x_+ - x_-) \leq |w^T(x_+ - x_-)| \leq ||w||||x_+ - x_-||;$$

consequently, we get that

$$||w|| \geq \frac{2}{||x_+ - x_-||}.$$

Since we seek to minimize $||w||$, we choose w^* such that

$$||w^*|| = \frac{2}{||x_+ - x_-||}.$$

In this case, as we want w^* to satisfy both constraints, we may note that

$$2 \leq w^{*T}(x_+ - x_-) \leq |w^{*T}(x_+ - x_-)| \leq ||w^*||||x_+ - x_-|| = 2.$$

This means that

$$|w^{*T}(x_+ - x_-)| = ||w^*||||x_+ - x_-||,$$

which can only happen when $w^* = k(x_+ - x_-)$. Since, we have already established that

$$||w^*|| = \frac{2}{||x_+ - x_-||},$$

we choose k to be

$$k = \frac{2}{||x_+ - x_-||^2}.$$

Now, we may write that

$$w^* = \frac{2(x_+ - x_-)}{||x_+ - x_-||^2}.$$

It remains to determine the value of b^*. To do that we fix the following equality

$$2\left(\frac{(x_+ - x_-)}{||x_+ - x_-||^2}\right)^T x_+ + b^* = 1;$$

which gives us that

$$\begin{aligned} b^* &= 1 - 2\frac{x_+^T x_+ - x_-^T x_+}{||x_+ - x_-||^2} \\ &= \frac{x_-^T x_- - x_+^T x_+}{||x_+ - x_-||^2} \\ &= \frac{||x_-||^2 - ||x_+||^2}{||x_+ - x_-||^2}. \end{aligned}$$

It is now easy to verify that (w^*, b^*) satisfies both constraints and minimizes $||w||$, and therefore gives us the optimal hyperplane.

Problem 8.2

In this case, the constraints are

$$-b \geq 1,\ -(-w_2 + b) \geq 1,\ (-2w_1 + b) \geq 1.$$

If we combine the first and the third ones, we get $w_1 \leq -1$. The quantity we seek to minimize is

$$\frac{1}{2}w^T w = \frac{1}{2}(w_1^2 + w_2^2) \geq \frac{1}{2}(1 + 0) \geq \frac{1}{2},$$

where we have equality when $w_1 = -1$ and $w_2 = 0$; consequently, we choose $w^* = (-1, 0)$. With this in mind, the third constraint becomes

$$1 \leq -2w_1^* + b = 2 + b \Leftrightarrow b \geq -1;$$

so we choose $b^* = -1$. It is now easy to verify that (w^*, b^*) satisfies both constraints and minimizes $||w||$, and therefore gives us the optimal hyperplane. The margin in this case is given by $1/||w^*|| = 1$.

Problem 8.3

(a) We begin by computing the Lagrangian, we get

$$\begin{aligned} \mathcal{L}(\alpha) &= \frac{1}{2}\sum_n \sum_m y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_n \alpha_n \\ &= \frac{1}{2}(8\alpha_2^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 - 4\alpha_2\alpha_3 + 4\alpha_3^2 + 6\alpha_3\alpha_4 - 6\alpha_4\alpha_2 + 6\alpha_3\alpha_4 + 9\alpha_4^2) - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4 \\ &= 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4. \end{aligned}$$

Concerning the constraints, we have that

$$0 = \sum_n y_n \alpha_n = -\alpha_1 - \alpha_2 + \alpha_3 + \alpha_4,$$

or equivalently

$$\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4$$

with $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0$.

(b) If we replace α_1 with $\alpha_3 + \alpha_4 - \alpha_2$, we obtain

$$\mathcal{L}(\alpha) = 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4.$$

(c) Now, we fix α_3 and α_4 and we take the derivative of $\mathcal{L}(\alpha)$ with respect to α_2, this gives us that

$$\frac{\partial \mathcal{L}}{\partial \alpha_2} = 8\alpha_2 - 4\alpha_3 - 6\alpha_4.$$

By setting the previous expression to 0, we get that

$$\alpha_2 = \frac{\alpha_3}{2} + \frac{3\alpha_4}{4},$$

and also that

$$\alpha_1 = -\alpha_2 + \alpha_3 + \alpha_4 = \frac{\alpha_3}{2} + \frac{\alpha_4}{4}.$$

These expressions are valid since they are both greater or equal to 0, and obviously

$$\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4.$$

(d) It remains to replace α_2 by its new expression (in (c)), we get that

$$\begin{aligned}
\mathcal{L}(\alpha) &= 4\left(\frac{\alpha_3}{2} + \frac{3\alpha_4}{4}\right)^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\left(\frac{\alpha_3}{2} + \frac{3\alpha_4}{4}\right)\alpha_3 - 6\left(\frac{\alpha_3}{2} + \frac{3\alpha_4}{4}\right)\alpha_4 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4 \\
&= \alpha_3^2 + (3\alpha_4 - 2)\alpha_3 + \frac{9}{4}\alpha_4^2 - 2\alpha_4 \\
&= \left(\alpha_3 + \frac{3\alpha_4 - 2}{2}\right)^2 + \frac{9}{4}\alpha_4^2 - 2\alpha_4 - \frac{(3\alpha_4 - 2)^2}{4} \\
&= \left(\alpha_3 + \frac{3\alpha_4 - 2}{2}\right)^2 + \alpha_4 - 1 \geq -1.
\end{aligned}$$

The minimum of the Lagrangian is attained when $\alpha_3 = 1$ and $\alpha_4 = 0$, in this case we also have

$$\alpha_1 = \frac{\alpha_3}{2} + \frac{\alpha_4}{4} = \frac{1}{2}$$

and

$$\alpha_2 = \frac{\alpha_3}{2} + \frac{3\alpha_4}{4} = \frac{1}{2}.$$

Problem 8.4

We have

$$X = \begin{pmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \end{pmatrix} \text{ and } y = \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}.$$

The Lagrangian is equal to

$$\begin{aligned}
\mathcal{L}(\alpha) &= \frac{1}{2}\sum_n \sum_m y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_n \alpha_n \\
&= 4\alpha_2^2 - 4\alpha_2\alpha_3 + 2\alpha_3^2 - \alpha_1 - \alpha_2 - \alpha_3 \\
&= 2(\alpha_1^2 - \alpha_1) + 2(\alpha_2^2 - \alpha_2) \geq -\frac{1}{2} - \frac{1}{2} \geq -1;
\end{aligned}$$

and the constraints are $\alpha_3 = \alpha_1 + \alpha_2$ with $\alpha_1, \alpha_2, \alpha_3 \geq 0$. The minimum of the Lagrangian is attained when $\alpha_1 = \alpha_2 = 1/2$ which gives us $\alpha_3 = \alpha_1 + \alpha_2 = 1$. Then, the optimal Lagrange multipliers are

$$\alpha_1^* = \frac{1}{2},\ \alpha_2^* = \frac{1}{2},\ \text{and } \alpha_3^* = 1.$$

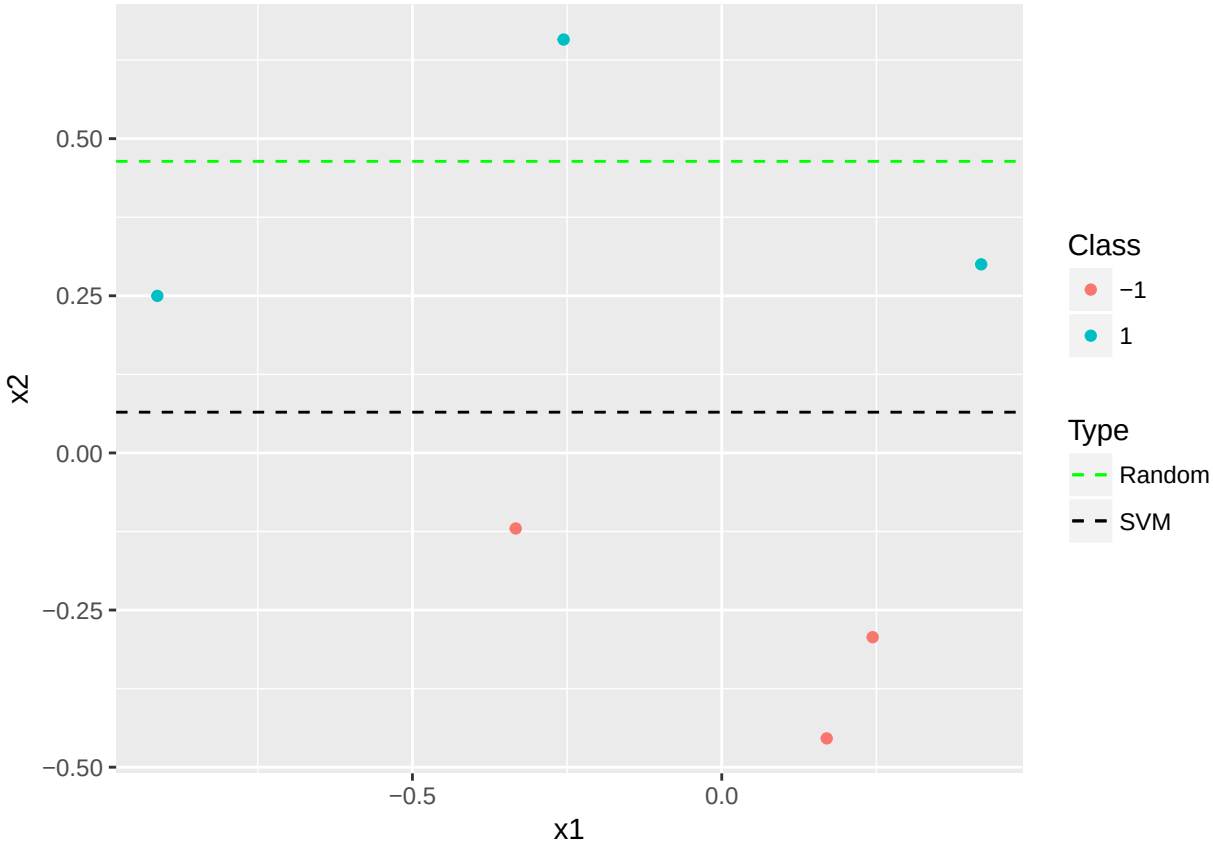
Problem 8.5

(a) Below, we generate three data points uniformly in the upper half of the input space and three data points in the lower half. We also obtain g_{random} and g_{SVM}.

```
set.seed(101)

f <- function(D) {

  return(sign(D$x2))
}

g <- function(D, a) {

  return(sign(D$x2 - a))
}

dataset_gen <- function() {
  D1 <- data.frame(x1 = runif(3, min = -1, max = 1), x2 = runif(3, min = 0, max = 1))
  D2 <- data.frame(x1 = runif(3, min = -1, max = 1), x2 = runif(3, min = -1, max = 0))
  D <- rbind(D1, D2)
  D <- cbind(D, y = as.factor(f(D)))

  return(D)
}

a_random <- function() {
  a <- runif(1, min = -1, max = 1)

  return(a)
}

a_SVM <- function(D) {
  min_pos <- min(D[D$x2 > 0, ]$x2)
  max_neg <- max(D[D$x2 < 0, ]$x2)
  a <- (min_pos + max_neg) / 2

  return(a)
}

D <- dataset_gen()

a_r <- a_random()
a_s <- a_SVM(D)
```

(b) Here, we create a plot of our data and of our two hypotheses.

```
ggplot(D, aes(x1, x2, col = y)) + geom_point() +
  guides(colour = guide_legend(title = "Class")) +
  geom_hline(aes(yintercept = a_r, linetype = "Random"), colour = "green") +
  geom_hline(aes(yintercept = a_s, linetype = "SVM"), colour = "black") +
  scale_linetype_manual(name = "Type", values = c(2, 2),
                        guide = guide_legend(override.aes = list(color = c("green", "black"))))
```

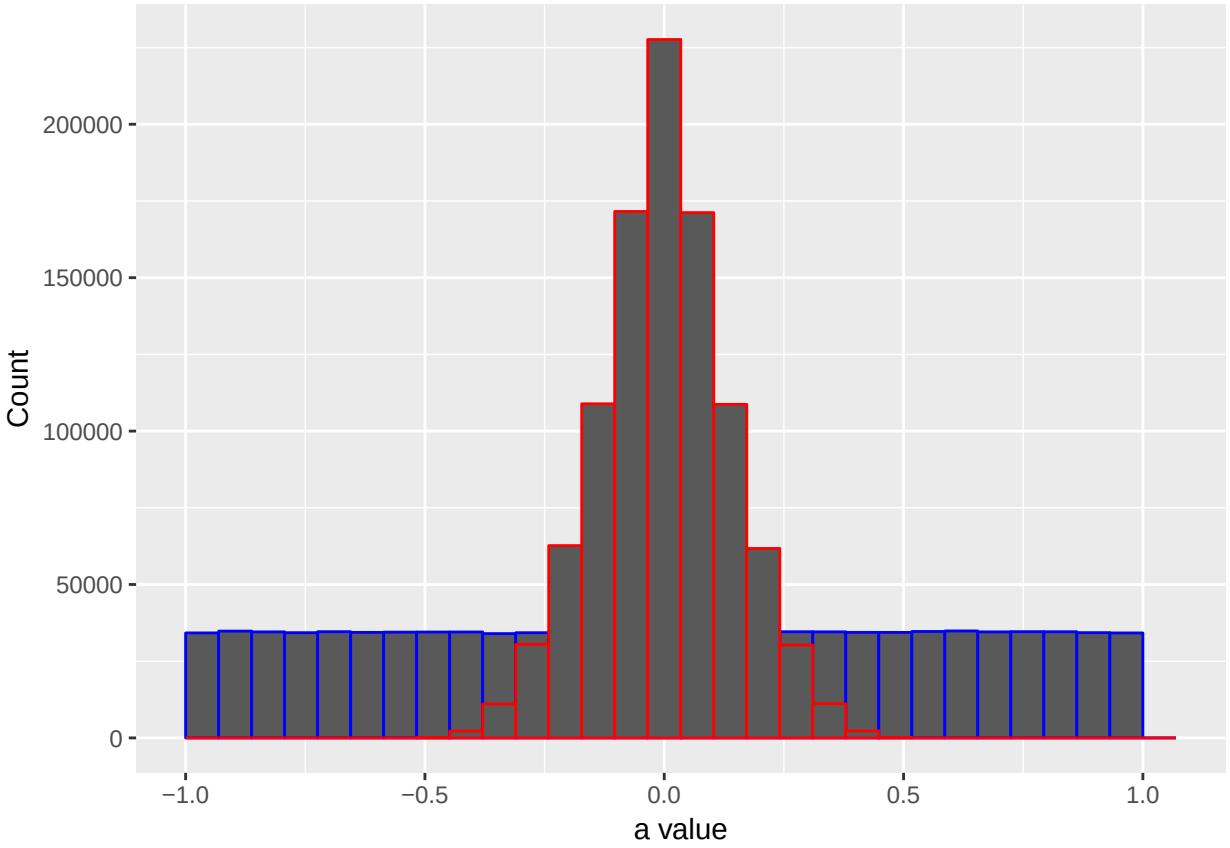


(*c*) Now, we repeat (*a*) for a million data sets to obtain one million random and SVM hypotheses.

```
N <- 1000000
vec_a_r <- rep(NA, N)
vec_a_s <- rep(NA, N)
for (i in 1:N) {
  D <- dataset_gen()
  a_r <- a_random()
  a_s <- a_SVM(D)
  vec_a_r[i] <- a_r
  vec_a_s[i] <- a_s
}

a_repeat <- data.frame(vec_a_r, vec_a_s)
```

(*d*) Below, we give a histogram of the values of a_{random} and another histogram of the values of a_{SVM}.

```
ggplot(a_repeat, aes(x = vec_a_r)) + geom_histogram(colour = "blue") +
  geom_histogram(aes(x = vec_a_s), colour = "red") +
  xlab("a value") + ylab("Count")
```



The two histograms are pretty different : the histogram of a_{random} is clearly uniform on $[-1, 1]$ which was expected; on the other hand, the histogram of a_{SVM} is symmetric and centered around 0 which was expected as well since our data set contains three points in the upper half and three points in the lower half, consequently the distribution of a_{SVM} was going to be centered around 0.

(e) Finally, we estimate the bias and variance for the two algorithms. Here, we cannot use any computation trick, so we have to use the full computation to get our estimates. We begin by estimating the biases of the two algorithms.

```
Ni <- 1000
Nj <- 1000

bias_r.x <- numeric(Nj)
bias_s.x <- numeric(Nj)
for (i in 1:Ni) {
  g_hat_r <- numeric(Ni)
  g_hat_s <- numeric(Ni)
  x.new <- data.frame(x1 = runif(1, min = -1, max = 1), x2 = runif(1, min = -1, max = 1))
  for (j in 1:Nj) {
    D <- dataset_gen()
    a_r <- a_random()
    a_s <- a_SVM(D)
    g_hat_r <- c(g_hat_r, g(x.new, a_r))
    g_hat_s <- c(g_hat_s, g(x.new, a_s))
  }
  g_bar_r <- mean(g_hat_r)
  bias_r.x <- c(bias_r.x, (g_bar_r - f(x.new))^2)
  g_bar_s <- mean(g_hat_s)
```

```
  bias_s.x <- c(bias_s.x, (g_bar_s - f(x.new))^2)
}
bias_r <- mean(bias_r.x)
bias_s <- mean(bias_s.x)
```

We get a bias of 0.2855502 for our random algorithm and a bias of 0.1571874 for SVM; the bias for SVM is clearly better than the bias for our random algorithm which is not surprising. Now, we take a look at the variances.

```
Ni <- 100
Nj <- 100

var_r.x <- numeric(Nj)
var_s.x <- numeric(Nj)
for (i in 1:Ni) {
  g_hat_r <- numeric(Ni)
  diff_hat_r <- numeric(Ni)
  g_hat_s <- numeric(Ni)
  diff_hat_s <- numeric(Ni)
  x.new <- data.frame(x1 = runif(1, min = -1, max = 1), x2 = runif(1, min = -1, max = 1))
  for (j in 1:Nj) {
    D <- dataset_gen()
    a_r <- a_random()
    g_hat_r <- c(g_hat_r, g(x.new, a_r))
    a_s <- a_SVM(D)
    g_hat_s <- c(g_hat_s, g(x.new, a_s))
  }
  g_bar_r <- mean(g_hat_r)
  g_bar_s <- mean(g_hat_s)
  for (j in 1:Nj) {
    D <- dataset_gen()
    a_r <- a_random()
    diff_hat_r <- c(diff_hat_r, (g(x.new, a_r) - g_bar_r)^2)
    a_s <- a_SVM(D)
    diff_hat_s <- c(diff_hat_s, (g(x.new, a_s) - g_bar_s)^2)
  }
  var_r.x <- c(var_r.x, mean(diff_hat_r))
  var_s.x <- c(var_s.x, mean(diff_hat_s))
}
var_r <- mean(var_r.x)
var_s <- mean(var_s.x)
```

We get a variance of 0.1884158 for our random algorithm and a variance of 0.0929888 for SVM; here the variance for SVM is clearly better as well. So, it seems that SVM is much better in our present case.

Problem 8.6

To find the stationary point of $\sum_{n=1}^{N} ||x_n - \mu||^2$, we compute its gradient and we equal it to 0, we get

$$\begin{aligned}\nabla_\mu \sum_{n=1}^{N} ||x_n - \mu||^2 &= \nabla_\mu \sum_{n=1}^{N} (x_n - \mu)^T (x_n - \mu) \\ &= \nabla_\mu \left(\sum_{n=1}^{N} (x_n^T x_n - 2x_n^T \mu + \mu^T \mu) \right) \\ &= 2N\mu - 2\sum_{n=1}^{N} x_n = 0.\end{aligned}$$

This gives us

$$\mu = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

To check that it is actually a minimum, we need to take a look at the Hessian matrix which is

$$Hess(\mu) = 2NI_N;$$

this is a positive definite matrix which means that our stationary point is in fact a minimum.

Problem 8.7

(a) We have that

$$\left\| \sum_{n=1}^{N} y_n x_n \right\|^2 = \left(\sum_{n=1}^{N} y_n x_n \right)^T \left(\sum_{m=1}^{N} y_m x_m \right) = \sum_{m,n} y_n y_m x_n^T x_m.$$

(b) Since $y_n \in \{-1, 1\}$, when $n = m$, we obviously have $y_n y_m = +1$ which gives us immediately $\mathbb{E}[y_n y_m] = 1$ when $n = m$. Now, if $n \neq m$, we may write that

$$\begin{aligned}\mathbb{P}[y_n y_m = 1] &= \mathbb{P}[y_n = 1, y_m = 1] + \mathbb{P}[y_n = -1, y_m = -1] \\ &= \frac{N/2}{N} \cdot \frac{N/2 - 1}{N - 1} + \frac{N/2}{N} \cdot \frac{N/2 - 1}{N - 1} \\ &= \frac{N - 2}{2(N - 1)}.\end{aligned}$$

In this case, we have that

$$\begin{aligned}\mathbb{E}[y_n y_m] &= (+1)\mathbb{P}[y_n y_m = 1] + (-1)\mathbb{P}[y_n y_m = -1] \\ &= \frac{N - 2}{2(N - 1)} - \left(1 - \frac{N - 2}{2(N - 1)} \right) \\ &= \frac{-1}{N - 1}.\end{aligned}$$

Consequently, we get

$$\mathbb{E}[y_n y_m] = \begin{cases} 1 \text{ if } n = m \\ \frac{-1}{N-1} \text{ if } n \neq m \end{cases}.$$

(c) We may write that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{n=1}^{N} y_n x_n\right\|^2\right] &= \mathbb{E}\left[\sum_{m,n} y_n y_m x_n^T x_m\right] \\
&= \sum_{m,n} \mathbb{E}[y_n y_m] x_n^T x_m \\
&= \sum_n \underbrace{\mathbb{E}[y_n y_m]}_{=1} x_n^T x_m + \sum_{m \neq n} \underbrace{\mathbb{E}[y_n y_m]}_{=-1/(N-1)} x_n^T x_m \\
&= \frac{N}{N-1} \sum_n x_n^T x_m - \underbrace{(\frac{1}{N-1} \sum_n x_n^T x_n + \frac{1}{N-1} \sum_{m \neq n} x_n^T x_m)}_{=1/(N-1)\sum_{m,n} x_n^T x_m} \\
&= \frac{N}{N-1} \sum_n x_n^T x_m - \frac{1}{N-1} \sum_{m,n} x_n^T x_m.
\end{aligned}$$

We may also write that

$$\begin{aligned}
\frac{N}{N-1} \sum_{n=1}^{N} ||x_n - \overline{x}||^2 &= \frac{N}{N-1} \sum_n (x_n^T x_n - 2 x_n^T \overline{x} + \overline{x}^T \overline{x}) \\
&= \frac{N}{N-1} \left(\sum_n x_n^T x_n - \underbrace{N \overline{x}^T \overline{x}}_{=(N/N^2)\sum_{m,n} x_n^T x_m} \right) \\
&= \frac{N}{N-1} \sum_n x_n^T x_n - \frac{1}{N-1} \sum_{m,n} x_n^T x_m.
\end{aligned}$$

This proves that

$$\mathbb{E}\left[\left\|\sum_{n=1}^{N} y_n x_n\right\|^2\right] = \frac{N}{N-1} \sum_{n=1}^{N} ||x_n - \overline{x}||^2.$$

(d) The Problem 8.6 gives us that

$$\sum_n ||x_n - \overline{x}||^2 \leq \sum_n ||x_n - \mu||^2$$

for all μ; so we get that

$$\sum_n ||x_n - \overline{x}||^2 \leq \sum_n \underbrace{||x_n||^2}_{\leq R^2} \leq NR^2.$$

(e) From parts (b) and (c), we have that

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{n=1}^{N} y_n x_n\right\|^2\right] &= \frac{N}{N-1} \sum_{n=1}^{N} ||x_n - \overline{x}||^2 \\
&\leq \frac{N}{N-1} NR^2 = \frac{N^2 R^2}{N-1}.
\end{aligned}$$

Now, let us assume that

$$\mathbb{P}\left[\left\|\sum_n y_n x_n\right\| \leq \frac{NR}{\sqrt{N-1}}\right] = 0,$$

or equivalently

$$\mathbb{P}\left[\left\|\sum_n y_n x_n\right\| > \frac{NR}{\sqrt{N-1}}\right] = 1.$$

In this case, we may write that

$$\underbrace{\mathbb{E}\left[\left\|\sum_{n=1}^N y_n x_n\right\|^2\right]}_{>N^2R^2/(N-1)} > \frac{N^2R^2}{N-1},$$

which is impossible. We conclude by writing that

$$\mathbb{P}\left[\left\|\sum_n y_n x_n\right\| \leq \frac{NR}{\sqrt{N-1}}\right] > 0.$$