Solutions by **Andrew Lys** `andrewlys(at)u.e.`

1. **Feature Selection.**

(a) i. Let $k = 1$. Then since x_1 and x_{100} are uncorrelated, we simply pick our feature as the feature which contributes more to the signal, namely x_100. Our predictor is then

$$h_w(x) = ax_{100}$$

And we pick $a = \frac{3}{\sqrt{10}}$. Our error is then:

$$L_D(h_w) = E[x_1^2/10] = \frac{1}{10}\operatorname{Var}(x_1) = \frac{1}{10}$$

For $k = 2$ and above, we simply pick x_1 and x_{100} as our features, $w = \frac{3}{\sqrt{10}}e_{100} + \frac{1}{\sqrt{10}}e_1$. We get zero loss. Therefore, we need $k = 2$ to get loss less than 0.01.

ii. For $k = 1$ we get the same predictor as the optimal feature selection.
For $k = 2$ we would select x_{100} first and the select x_1 as the second feature.
For $k > 2$, we would only select x_{100} and x_1 as our features, since adding any feature would increase our loss.
Therefore, the optimal feature selection is the same as the feature selection with $k = 2$, and we require $k = 2$ to get loss less than 0.01.

iii. We solve the following optimization problem for a fixed B.

$$\arg\min_w E\left(\frac{3}{\sqrt{10}}x_{100} + \frac{1}{\sqrt{10}}x_1 - \langle w, x\rangle\right)^2 \qquad \text{s.t.} \quad \|w\|_1 \leq B$$

Where the expectation is taken over the distribution of x. It is clear that any non-zero coefficient on w_i, for $i \neq 1, 100$ is inefficient, since it pointlessly increases our loss and increases our ℓ^1 norm. Therefore, the support of w is at most $\{1, 100\}$, for all values of B, and we can solve the optimization problem by solving the following optimization problem:

$$\arg\min_{w_1, w_{100}} E\left(\left(\frac{3}{\sqrt{10}} - w_{100}\right)x_{100} + \left(\frac{1}{\sqrt{10}} - w_1\right)x_1\right)^2 \qquad \text{s.t.} \quad |w_1| + |w_{100}| \leq B$$

We can expand the objective function as follows:

$$\begin{aligned}
E(\ell) &= E\left(\left(\frac{3}{\sqrt{10}} - w_{100}\right)x_{100} + \left(\frac{1}{\sqrt{10}} - w_1\right)x_1\right)^2 \\
&= E\left(\left(\frac{3}{\sqrt{10}} - w_{100}\right)^2 x_{100}^2 + \left(\frac{1}{\sqrt{10}} - w_1\right)^2 x_1^2 + 2\left(\frac{3}{\sqrt{10}} - w_{100}\right)\left(\frac{1}{\sqrt{10}} - w_1\right)x_1x_{100}\right) \\
&= \left(\frac{3}{\sqrt{10}} - w_{100}\right)^2 E[x_{100}^2] + \left(\frac{1}{\sqrt{10}} - w_1\right)^2 E[x_1^2] + 2\left(\frac{3}{\sqrt{10}} - w_{100}\right)\left(\frac{1}{\sqrt{10}} - w_1\right)E[x_1x_{100}] \\
&= \left(\frac{3}{\sqrt{10}} - w_{100}\right)^2 + \left(\frac{1}{\sqrt{10}} - w_1\right)^2 + 2\left(\frac{3}{\sqrt{10}} - w_{100}\right)\left(\frac{1}{\sqrt{10}} - w_1\right)\cdot 0 \\
&= \left(\frac{3}{\sqrt{10}} - w_{100}\right)^2 + \left(\frac{1}{\sqrt{10}} - w_1\right)^2
\end{aligned}$$

From this it is clear that w_1 and w_{100} are always positive. Additionally, for $B \leq \frac{4}{\sqrt{10}}$, the constraint is active, so we can write our constraints as:

$$\begin{aligned}
w_1 &\geq 0 \\
w_{100} &\geq 0 \\
w_1 + w_{100} &= B
\end{aligned}$$

For the case where the optimal solution is on the interior of the constraint, we solve the following system of equations:

$$w_1 + w_{100} = B$$

$$w_1 = w_{100} - \frac{2}{\sqrt{10}}$$

This gives us the solution:

$$w_1 = \frac{B}{2} - \frac{1}{\sqrt{10}}$$
$$w_{100} = \frac{B}{2} + \frac{1}{\sqrt{10}}$$

When the solution is an extreme point, we have the solution:

$$w_1 = 0$$
$$w_{100} = B$$

And this occurs when

$$B \leq \frac{2}{\sqrt{10}}$$

Therefore, when $B \leq \frac{2}{\sqrt{10}}$, we have $k = 1$, and when $B > \frac{2}{\sqrt{10}}$, we have $k = 2$. For $k = 1$, we have the same as the optimal solution, namely $w = \frac{3}{\sqrt{10}} e_{100}$, and the loss is $\frac{1}{10}$. For $\frac{2}{\sqrt{10}} \leq B \leq \frac{4}{\sqrt{10}}$, we have:

$$w = \left(\frac{B}{2} + \frac{1}{\sqrt{10}}\right) e_{100} + \left(\frac{B}{2} - \frac{1}{\sqrt{10}}\right) e_1$$

And the loss is:

$$\begin{aligned} L_D(h_w) &= \left(\frac{3}{\sqrt{10}} - \frac{B}{2} - \frac{1}{\sqrt{10}}\right)^2 + \left(\frac{1}{\sqrt{10}} - \frac{B}{2} + \frac{1}{\sqrt{10}}\right)^2 \\ &= 2\left(\frac{B}{2} - \frac{2}{\sqrt{10}}\right)^2 \end{aligned}$$

When $B > \frac{4}{\sqrt{10}}$, we have $k = 2$, and the loss is zero.

iv. We calculate the correlation coefficient for each x_i and y. For $i = 1$, we have:

$$\begin{aligned} \rho_1 &= \frac{E[x_1 y]}{\sqrt{E[x_1^2] E[(y - E(y))^2]}} \\ &= \frac{E\left[\frac{1}{\sqrt{10}} x_1^2 + \frac{3}{\sqrt{10}} x_{100} x_1\right]}{\sqrt{\operatorname{Var}(x_1)\operatorname{Var}(y)}} \\ &= \frac{\frac{1}{\sqrt{10}} E[x_1^2] + \frac{3}{\sqrt{10}} E[x_{100}] E[x_1]}{\sqrt{\operatorname{Var}\left(\frac{1}{\sqrt{10}} x_1 + \frac{3}{\sqrt{10}} x_{100}\right)}} \\ &= \frac{\frac{1}{\sqrt{10}}}{\sqrt{\frac{1}{10} + \frac{9}{10}}} \\ &= \frac{1}{\sqrt{10}} \end{aligned}$$

Note that for each i, the denominator is the same, 1.
For $i = 100$, we have:

$$\begin{aligned} \rho_{100} &= \frac{E[x_{100} y]}{\sqrt{E[x_{100}^2] E[(y - E(y))^2]}} \\ &= E\left[\frac{1}{\sqrt{10}} x_1 x_{100} + \frac{3}{\sqrt{10}} x_{100}^2\right] \\ &= \frac{3}{\sqrt{10}} \end{aligned}$$

for $i \neq 1, 100$, we have:

$$\begin{aligned}\rho_i &= \frac{E[x_i y]}{\sqrt{E[x_i^2]E[(y-E(y))^2]}} \\ &= E\left[\frac{1}{\sqrt{10}}x_i x_1 + \frac{3}{\sqrt{10}}x_i x_{100}\right] \\ &= \frac{3}{\sqrt{10}}E[x_i x_{100}] \\ &= \frac{3}{\sqrt{10}} \cdot \frac{9}{10} = \frac{2.7}{\sqrt{10}}\end{aligned}$$

Therefore, if $k = 1$, we select x_{100} as our feature, which is the ideal feature. If $k = 2$, we select x_{100} and any other feature different from x_1. For $k = 3, \dots, 99$, we select any features different from x_1. Only for $k = 100$, do we finally select x_1.

For $k < 100$, we have the same loss as if we only selected x_{100}, since the coefficient of the features we select, other than x_{100}, is zero. We thus get loss $\frac{1}{10}$, and only for $k = 100$ do we get loss 0.

(b) i. Let $k = 1$. Then we compute the loss for x_1, x_2, and x_i, for $i \neq 1, 2$ as the chosen features.

$$\begin{aligned}E[(y - az_1)^2] = E[(z_2 - az_1)^2] &= E[(z_1 - az_2)^2] - E[z_1 - az_2]^2 = \mathrm{Var}[z_1 - az_2] \\ &= \mathrm{Var}[z_1] + a^2\,\mathrm{Var}[z_2] = 1 + a^2 \\ \implies a = 0, &\qquad L_D(h_{w_1}) = 1\end{aligned}$$

For $k = 2$, we replace the coefficient of z_2 with $b = 0.0001$, so we can reuse the same calculations for z_i.

$$\begin{aligned}E[(y - az_1 - abz_2)^2] &= E[(z_2 - az_1 - abz_2)^2] = \mathrm{Var}(z_2 - az_1 - abz_2) \\ &= (1-ab)^2 + a^2 \\ \frac{\partial L}{\partial a}&(1-ab)^2 + a^2 = 0 \\ 0 &= 2(1-ab)(-b) + 2a \\ a &= b - ab^2 \\ b &= a(1+b^2) \\ a &= \frac{1}{b^{-1}+b} \\ \implies L &= \left(1 - \frac{b}{b^{-1}+b}\right)^2 + \left(\frac{b}{b^{-1}+b}\right)^2 \\ &= \frac{1}{b^{-2}+1}\end{aligned}$$

Therefore, selecting x_2 our feature we have that our loss is:

$$L_D(h_{w_2}) = \frac{1}{10^{-8}+1}$$

Since the z_i have no covariance between each other, we have the same logic for x_i. Selecting x_i for $i \neq 1, 2$, we have:

$$L_D(h_{w_i}) = \frac{1}{10^{-6}+1}$$

Therefore, for $k = 1$, we would select any x_i for $i \neq 1, 2$.

For $k = 2$, we would select x_1 and x_2 as our features, since we get zero loss by selecting $w = -10^4 e_1 + 10^4 e_2$. This gives us:

$$\begin{aligned}h_w(x) &= -10^4 x_1 + 10^4 x_2 = -10^4 z_1 + 10^4(z_1 + 10^{-4} z_2) \\ &= z_2\end{aligned}$$

Therefore, we have:

$$L_D(h_w) = E[(y - h_w(x))^2] = E[(z_2 - z_2)^2] = 0$$

Thus, for $k > 2$, we similarly select x_1 and x_2 as our features.

ii. For greedy feature selection we select x_i, $i > 2$ as our first feature, since they have the lowest loss. We now calculate the loss for $k = 2$, given that we've selected some x_i as our first feature.
We begin with x_1 as our first second feature.

$$\begin{aligned}\ell &= E[(z_2 - az_1 - bz_i - b10^{-3}z_2)^2] \\ &= \mathrm{Var}((1 - b10^{-3})z_2 - az_1 - bz_i) \\ &= (1 - b10^{-3})^2 + a^2 + b^2\end{aligned}$$

Since a^2 is non-negative, we set $a = 0$ to minimize the loss, and we choose b as in the optimal case for $k = 1$.
We check x_2 as our second feature.

$$\begin{aligned}\ell &= E[(z_2 - az_1 - a10^{-4}z_2 - bz_i - b10^{-3}z_2)^2] \\ &= \mathrm{Var}((1 - a10^{-4} - b10^{-3})z_2 - az_1 - bz_i) \\ &= (1 - a10^{-4} - b10^{-3})^2 + a^2 + b^2\end{aligned}$$

We check x_j where $j \neq i$ and $j > 2$ as our second feature. With the same logic, we have:

$$\ell = (1 - (a + b)10^{-3})^2 + a^2 + b^2$$

Clearly, for every choice of a and b, we have that the loss is minimized when we select x_j as our next feature. Continuing this logic, for $k = 3, \ldots, 98$, we select x_i where $i > 2$ as our features. Our loss is then:

$$\ell = \left(1 - \left(\sum_{i=1}^{k} a_{i+2}\right) 10^{-3}\right)^2 + \sum_{i=1}^{k} a_{i+2}^2$$

Taking the critical points, we have:

$$\begin{aligned}&[a_j] : -2\left(1 - 10^{-3}\sum_{i=1}^{k} a_{i+2}\right)10^{-3} + 2a_j = 0 \\ \implies\ & a_j = 10^{-3}\left(1 - 10^{-3}\sum_{i=1}^{k} a_{i+2}\right) \\ & a_j = 10^{-3} - 10^{-6}ka_j \\ & a_j = \frac{10^{-3}}{1 + 10^{-6}k}\end{aligned}$$

Where we have all the a_js are equal by the second line. Therefore, the loss is:

$$\begin{aligned}L_D(h_w) &= \left(1 - 10^{-3} \cdot k \cdot \frac{10^{-3}}{1 + 10^{-6}k}\right)^2 + k \cdot \left(\frac{10^{-3}}{1 + 10^{-6}k}\right)^2 \\ &= \left(1 - \frac{10^{-6}k}{1 + 10^{-6}k}\right)^2 + \frac{10^{-6}k}{(1 + 10^{-6}k)^2} \\ &= \frac{1}{(1 + 10^{-6}k)^2} + \frac{10^{-6}k}{(1 + 10^{-6}k)^2} \\ &= \frac{1 + 10^{-6}k}{(1 + 10^{-6}k)^2} \\ &= \frac{1}{1 + 10^{-6}k}\end{aligned}$$

For $k = 99$, we select x_2 as our feature, since adding x_1 wouldn't decrease our loss. We get the following loss:

$$\begin{aligned}
L_D(h_w) &= E\left[\left(z_2 - 10^{-3}\sum_{i=3}^{100} a_i z_2 - 10^{-4} a_2 z_2 + \sum_{i=3}^{100} a_i z_i + a_2 z_1\right)^2\right] \\
&= \mathrm{Var}\left(\left(1 - 10^{-3}\sum_{i=3}^{100} a_i - 10^{-4} a_2\right) z_2 + \sum_{i=3}^{100} a_i z_i + a_2 z_1\right) \\
&= \left(1 - 10^{-3}\sum_{i=3}^{100} a_i - 10^{-4} a_2\right)^2 + \sum_{i=2}^{100} a_i^2
\end{aligned}$$

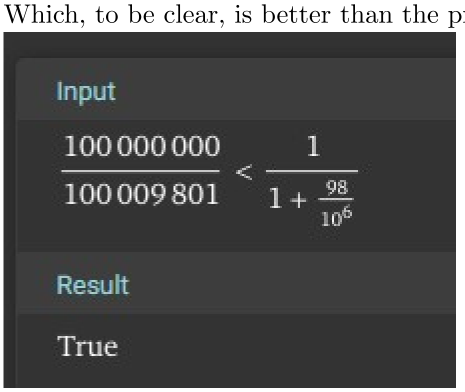
Taking critical points with a_2 and a_i, $i > 2$, we have:

$$\begin{aligned}
[a_i] : -2\left(1 - 10^{-3}\sum_{i=3}^{100} a_i - 10^{-4} a_2\right) 10^{-3} + 2a_i = 0 \\
\implies a_i = 10^{-3}\left(1 - 10^{-3}\sum_{i=3}^{100} a_i - 10^{-4} a_2\right) \\
[a_2] : -2\left(1 - 10^{-3}\sum_{i=3}^{100} a_i - 10^{-4} a_2\right) 10^{-4} + 2a_2 = 0 \\
\implies a_2 = 10^{-4}\left(1 - 10^{-3}\sum_{i=3}^{100} a_i - 10^{-4} a_2\right) \\
\implies a_2 = 10^{-4}\left(1 - 10^{-3} 98 a_i - 10^{-4} a_2\right) \\
\implies a_i = 10^{-3}\left(1 - 10^{-3} 98 a_i - 10^{-4} a_2\right) \\
\implies a_2 = \frac{10^5}{10^9 + 98 \cdot 10^2 + 1} \\
\implies a_i = \frac{10^6}{10^9 + 98 \cdot 10^2 + 1}
\end{aligned}$$

Plugging these values into our loss function, we get:

$$\begin{aligned}
L_D(h_w) &= \left(1 - 10^{-3} 98 \cdot a_i - 10^{-4} a_2\right)^2 + 98 \cdot a_i^2 + a_2^2 \\
&= 0.999901...
\end{aligned}$$

Which, to be clear, is better than the previous results (verified by wolfram alpha).



This is still (obviously) worse than the target of 0.01. Therefore, the only k for which we get loss lower than 0.01 is $k = 100$, in which we finally are able to select x_1 and x_2, and we let $w = -10^4 e_1 + 10^4 e_2$. As shown above, this achieves zero loss.

2. **Boosting as Coordinate Descent.**