Topics in ML: Homework #2

Due on Thursday 23rd, 2023 at 23:59pm

Jacopo Peroni

While solving this homework I discussed some of the problems with Marco Scialanga, Oskar Koiner, and Aryan Rahbari.

Problem 1

Subproblem 1.1

Before studying the minimizer of the empirical risk, it's useful to see that

$$\mathbf{X}^T = \left(\mathbf{X}^T\mathbf{X}\right)\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}$$

Now we can solve the subproblem by starting noticing that the function $\|\mathbf{X}\omega - \mathbf{y}\|_2^2$ is convex in ω and continuous so it has a minimizer, and that $\mathbf{X}\mathbf{X}^T$ is invertible because the $\text{rank}(\mathbf{X}) = n$ so the $\text{rank}(\mathbf{X}\mathbf{X}^T)$ is maximum.

$$\mathcal{R}(\omega) = \|\mathbf{X}\omega - \mathbf{y}\|_2^2 = (\mathbf{X}\omega - \mathbf{y})^T(\mathbf{X}\omega - \mathbf{y}) = \omega^T\mathbf{X}^T\mathbf{X}\omega - \omega^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{X}\omega + \mathbf{y}^T\mathbf{y}$$

To find a minimum let's take a look at $\nabla\mathcal{R}(\omega) = 0$

$$\nabla\mathcal{R}(\omega) = 2\mathbf{X}^T\mathbf{X}\omega - 2\mathbf{X}^T\mathbf{y} = 0 \implies \mathbf{X}^T\mathbf{X}\omega = \mathbf{X}^T\mathbf{y} \implies \mathbf{X}^T\mathbf{X}\omega = \left(\mathbf{X}^T\mathbf{X}\right)\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}$$

$$\implies \omega = \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y} + \omega^\perp$$

With $\omega^\perp \in \ker(\mathbf{X}^T\mathbf{X})$.
Noticing that from the formula above $\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y} \in \text{Im}(\mathbf{X}^T\mathbf{X})$,we know that

$$\|\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y} + \omega^\perp\|_2^2 = \|\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}\|_2^2 + \|\omega^\perp\|_2^2 \geq \|\mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}\|_2^2$$

Thus the euclidean norm empirical risk minimizer is

$$\hat{\omega} = \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\mathbf{y}$$

Subproblem 1.2

Before solving the subproblem it's important to notice that

$$Y_i = \langle \omega^*, X_i \rangle + Z_i$$

$$R(\hat{\omega}) = \mathbf{E}_{(X,Y)} \left[(\langle \hat{\omega}, X \rangle - Y)^2 \right]$$

$$R(\omega^*) = \mathbf{E}_{(X,Y)} \left[(\langle \omega^*, X \rangle - Y)^2 \right]$$

To solve the subproblem it's useful to expand the calculations

$$\mathbf{E}_{\mathbf{z}} \mathbf{E}_{(X,Y)} \left[(\langle \hat{\omega}, X \rangle - Y)^2 \right] - \mathbf{E}_{(X,Y)} \left[(\langle \omega^*, X \rangle - Y)^2 \right] =$$

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \mathbf{E}_Y \left[(\langle \hat{\omega}, X \rangle - Y)^2 \, | X = x \right] - \mathbf{E}_X \mathbf{E}_Y \left[(\langle \omega^*, X \rangle - Y)^2 \, | X = x \right] =$$

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \mathbf{E}_Y \left[\langle \hat{\omega}, X \rangle^2 - 2 \langle \hat{\omega}, X \rangle Y + Y^2 | X = x \right] - \mathbf{E}_X \mathbf{E}_Y \left[\langle \omega^*, X \rangle^2 - 2 \langle \omega^*, X \rangle Y + Y^2 | X = x \right] =$$

Now using that $\mathbf{E}_Y[\langle \omega_i, X \rangle] = \langle \omega_i, X \rangle$ for $\omega_i \in \{\hat{\omega}, \omega^*\}$ and the linearity of the expected value

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[\langle \hat{\omega}, X \rangle^2 - 2 \langle \hat{\omega}, X \rangle \mathbf{E}_Y \left[Y | X = x \right] + \mathbf{E}_Y \left[Y^2 | X = x \right] \right] +$$

$$- \mathbf{E}_X \left[\langle \omega^*, X \rangle^2 - 2 \langle \omega^*, X \rangle \mathbf{E}_Y \left[Y | X = x \right] + \mathbf{E}_Y \left[Y^2 | X = x \right] \right] =$$

Using that $Y|X = x \sim \langle \omega^*, X \rangle + Z$ with Z as a zero-mean random variable, that $\mathbf{E}_X[\mathbf{E}_Y[Y^2|X = x]] = \mathbf{E}_Y[Y^2]$ and that $\mathbf{E}_{\mathbf{z}}[\mathbf{E}_Y[Y^2|X = x]] = \mathbf{E}_Y[Y^2]$ and the linearity of the expected value

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[\langle \hat{\omega}, X \rangle^2 - 2 \langle \hat{\omega}, X \rangle \langle \omega^*, X \rangle \right] + \mathbf{E}_Y \left[Y^2 \right] - \mathbf{E}_X \left[\langle \omega^*, X \rangle^2 - 2 \langle \omega^*, X \rangle^2 \right] - \mathbf{E}_Y \left[Y^2 \right] =$$

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[\langle \hat{\omega}, X \rangle^2 - 2 \langle \hat{\omega}, X \rangle \langle \omega^*, X \rangle \right] + \mathbf{E}_X \left[\langle \omega^*, X \rangle^2 \right] =$$

Using that $\langle \omega^*, X \rangle^2$ doesn't depend on $\mathbf{z}$

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[\langle \hat{\omega}, X \rangle^2 - 2 \langle \hat{\omega}, X \rangle \langle \omega^*, X \rangle \right] + \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[\langle \omega^*, X \rangle^2 \right] =$$

Using the linearity of the expected value and of the inner product we get

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[(\langle \hat{\omega}, X \rangle - \langle \omega^*, X \rangle)^2 \right] = \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[(\langle \hat{\omega} - \omega^*, X \rangle)^2 \right] =$$

From the result of subproblem 1.1

$$= \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[\left(\langle \mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} \mathbf{y} - \omega^*, X \rangle \right)^2 \right] = \mathbf{E}_{\mathbf{z}} \mathbf{E}_X \left[\left(\langle \mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} (\mathbf{X} \omega^* + \mathbf{z}) - \omega^*, X \rangle \right)^2 \right] =$$

Let's focus on the term inside the expected values

$$\left(\langle \mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} (\mathbf{X} \omega^* + \mathbf{z}) - \omega^*, X \rangle \right)^2 =$$

$$= \left(\langle \mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} (\mathbf{X} \omega^* + \mathbf{z}) - \omega^*, X \rangle \right) \left(\langle X, \mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} (\mathbf{X} \omega^* + \mathbf{z}) - \omega^* \rangle \right) =$$

$$= \left(\mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} \mathbf{X} \omega^* + \mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} \mathbf{z} - \omega^* \right)^T X X^T \left(\mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} \mathbf{X} \omega^* + \mathbf{X}^T \left(\mathbf{X} \mathbf{X}^T \right)^{-1} \mathbf{z} - \omega^* \right) =$$

$$= \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right)^T XX^T \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right) +$$
$$+ \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right)^T XX^T \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right) +$$
$$+ \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right)^T XX^T \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right) +$$
$$+ \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right)^T XX^T \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right) =$$

Now using the linearity and the homogeneity of the expected value respect to X this expression becomes

$$= \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right)^T \Sigma \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right) +$$
$$+ \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right)^T \Sigma \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right) +$$
$$+ \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right)^T \Sigma \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right) +$$
$$+ \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right)^T \Sigma \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{z}\right) =$$

It's important to notice that the second and the third terms are linearly dependent on $\mathbf{z}$, so computing the $\mathbf{E}_\mathbf{z}$ cancels them (because $\mathbf{z}$ is a vector of zero mean variables), so in the end the difference of the risks becomes

$$\left\|\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right\|_\Sigma^2 + \mathbf{E}_\mathbf{z}\left[\mathbf{z}^T \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)^T \Sigma \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)\mathbf{z}\right] =$$

$$= \left\|\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right\|_\Sigma^2 + \mathbf{E}_\mathbf{z}\left[\mathrm{Tr}\left(\mathbf{z}^T \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)^T \Sigma \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)\mathbf{z}\right)\right] =$$

Using the cyclic property of the trace we obtain

$$= \left\|\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right\|_\Sigma^2 + \mathbf{E}_\mathbf{z}\left[\mathrm{Tr}\left(\mathbf{z}\mathbf{z}^T \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)^T \Sigma \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)\right)\right] =$$

To conclude we use the linearity of the trace operator and that $((\mathbf{X}\mathbf{X}^T)^{-1})^T = (\mathbf{X}\mathbf{X}^T)^{-1}$ so we get

$$= \left\|\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right\|_\Sigma^2 + \mathrm{Tr}\left(\mathbf{E}_\mathbf{z}\left[\mathbf{z}\mathbf{z}^T\right]\left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\Sigma\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right) =$$

$$= \left\|\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\omega^*\right\|_\Sigma^2 + \sigma^2 \mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\Sigma\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)$$

Subproblem 1.3

To prove the equivalence between the two terms we expand the RHS

$$\left\| \left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right\|^2_{\Sigma - \frac{1}{n}\mathbf{X}^T\mathbf{X}} =$$

$$= \left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right)^T \left(\Sigma - \frac{1}{n}\mathbf{X}^T\mathbf{X} \right) \left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right) =$$

Using the definition of B^2

$$= B^2 - \frac{1}{n} \left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right)^T \mathbf{X}^T\mathbf{X} \left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right)$$

If we prove that the second term is equal to zero we have proved the equivalence

$$\left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right)^T \mathbf{X}^T\mathbf{X} \left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right) =$$

$$= {\omega^*}^T \left[\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} \right)^T - I_d \right] \mathbf{X}^T\mathbf{X} \left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* =$$

Using that $(\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X})^T = ((\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X})^T\mathbf{X} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$

$$= {\omega^*}^T\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\omega^* - {\omega^*}^T\mathbf{X}^T\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\omega^* +$$

$$-{\omega^*}^T\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{X}^T\mathbf{X}\omega^* + {\omega^*}^T\mathbf{X}^T\mathbf{X}\omega^* =$$

$$= {\omega^*}^T\mathbf{X}^T\mathbf{X}\omega^* - {\omega^*}^T\mathbf{X}^T\mathbf{X}\omega^* - {\omega^*}^T\mathbf{X}^T\mathbf{X}\omega^* + {\omega^*}^T\mathbf{X}^T\mathbf{X}\omega^* = 0$$

To prove the inequality we can see that, given the equality just proven, we have

$$B^2 = \left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right)^T \left(\Sigma - \frac{1}{n}\mathbf{X}^T\mathbf{X} \right) \left(\left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right) \leq$$

$$\leq \left\| \Sigma - \frac{1}{n}\mathbf{X}^T\mathbf{X} \right\| \left\| \left[\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right] \omega^* \right\|_2^2 \leq$$

$$\leq \left\| \Sigma - \frac{1}{n}\mathbf{X}^T\mathbf{X} \right\| \left\| \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right\|_2^2 \|\omega^*\|_2^2$$

The problem becomes proving that $\left\| \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right\|_2^2 \leq 1$, to do that we notice a fundamental property of the matrix

$$\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right)^T \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right) =$$

$$= \left(\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} \right)^T - I_d \right) \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right) =$$

$$= \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right) \left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T \right)^{-1} \mathbf{X} - I_d \right) =$$

$$= \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - 2\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} + I_d = -\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)$$

So $\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d$ is a matrix of the type $A^2 = -A$.
If we look at what are the possible eigenvalues we get that

$$-\lambda v = -Av = A^2 v = A\lambda v = \lambda A v = \lambda^2 v \Longrightarrow \left(\lambda^2 + \lambda\right) v = 0 \Longrightarrow \lambda \in \{0, -1\}$$

Thus

$$\left\|\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right\|_2^2 = \left(\lambda_{max}\left(\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} - I_d\right)\right)^2 \leq 1$$

So we get the conclusion.

Subproblem 1.4

If we assume that $X_i \sim N(0, I_d)$ for all i we have that in the definition of Σ we have that $X \sim N(0, I_d)$ as well. This means that Σ is just the covariance matrix of this normal distribution, so $\Sigma = I_d$.
Given this, the variance becomes

$$V = \mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X}\mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right) = \mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)$$

To tackle the computation of $\mathbf{E}[V]$ we use the linear property of the trace

$$\mathbf{E}\left[\mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)\right] = \mathrm{Tr}\left(\mathbf{E}\left[\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right]\right)$$

Using the hint, if we consider as U_i the columns of the matrix $\mathbf{X}$ we have that they are $\sim N(0, I_n)$ and are i.i.d.
From a simple computation, we can see that

$$\left(U_i U_i^T\right)_{j,k} = \mathbf{X}_{j,i}\mathbf{X}^T_{i,k} \Longrightarrow \left(\sum_{i=1}^n U_i U_i^T\right)_{j,k} = \sum_{i=1}^n \mathbf{X}_{j,i}\mathbf{X}^T_{i,k} = \left(\mathbf{X}\mathbf{X}^T\right)_{j,k}$$

Thus

$$\mathrm{Tr}\left(\mathbf{E}\left[\left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right]\right) = \mathrm{Tr}\left(\mathbf{E}\left[\left(\sum_{i=1}^n U_i U_i^T\right)^{-1}\right]\right) = \mathrm{Tr}\left(\frac{1}{d-n-1} I_n\right) = \frac{n}{d-n-1}$$

In the second passage, we used the hint (note that we know that $d > n$ but not that $d > n+1$ as required by the hint, let's set $d > n+1$ given that later we will have $d \gg n$).
Looking at the computations in the subproblem 1.2, where we derived the equivalence

$$\mathbf{E}_{\mathbf{z}} R\left(\hat{\omega}\right) - R\left(\omega^*\right) = B^2 + \sigma^2 V$$

It's clear that $R\left(\omega^*\right)$ isn't dependent on $\mathbf{X}$ so

$$\mathbf{E}_{\mathbf{X},\mathbf{z}} R\left(\hat{\omega}\right) - R\left(\omega^*\right) = \mathbf{E}_{\mathbf{X}}\left[B^2 + \sigma^2 V\right] = \mathbf{E}_{\mathbf{X}}\left[B^2\right] + \sigma^2 \mathbf{E}_{\mathbf{X}}\left[V\right]$$

As stated in the problem we consider the case $d \gg n$, that imply $d > n + 1$, noticing that B^2 is a constant, we have

$$\mathbf{E}_{\mathbf{X}}\left[B^2\right] + \sigma^2 \mathbf{E}_{\mathbf{X}}\left[V\right] = B^2 + \frac{\sigma^2 n}{d - n - 1} \approx B^2$$

To upper bound B^2 we can use the result from subproblem 1.3. On top of that, we can also use the remark, in fact, the rows of $\mathbf{X}$ are normally distributed with zero mean, so they are sub-Gaussian with zero mean and $\mathrm{Tr}(\Sigma) = d$ so $\mathrm{Tr}(\Sigma)/\|\Sigma\| \leq d$.
This ensures that the expected risk $\mathbf{E}_{\mathbf{X},\mathbf{z}} R\left(\hat{\omega}\right) - R\left(\omega^*\right)$ is small.

Subproblem 1.5

Now changing the distribution to

$$X \sim N\left(0, \begin{bmatrix} I_k & 0 \\ 0 & \epsilon I_{d-k} \end{bmatrix}\right)$$

We get that Σ is still the covariance matrix

$$\Sigma = \begin{bmatrix} I_k & 0 \\ 0 & \epsilon I_{d-k} \end{bmatrix}$$

So if we insert this into the definition of V we get

$$V = \mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{X} \begin{bmatrix} I_k & 0 \\ 0 & \epsilon I_{d-k} \end{bmatrix} \mathbf{X}^T \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right)$$

Let's introduce the notation $A^{h,k}$ to indicate the matrix made of the columns from the h-th to the k-th of the matrix A. Analogously $A_{h,k}$ for the rows.
With this notation, we can rewrite $\mathbf{X}\Sigma\mathbf{X}^T$ as

$$\begin{bmatrix}\mathbf{X}^{1,k} & \mathbf{X}^{k+1,d}\end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & \epsilon I_{d-k} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{1,k}^T \\ \mathbf{X}_{k+1,n}^T \end{bmatrix} = \begin{bmatrix}\mathbf{X}^{1,k} & \epsilon\mathbf{X}^{k+1,d}\end{bmatrix} \begin{bmatrix} \mathbf{X}_{1,k}^T \\ \mathbf{X}_{k+1,d}^T \end{bmatrix} = \mathbf{X}^{1,k}\mathbf{X}_{1,k}^T + \epsilon\mathbf{X}^{k+1,d}\mathbf{X}_{k+1,d}^T =$$

$$= \mathbf{X}^{1,k}\left(\mathbf{X}^{1,k}\right)^T + \epsilon\mathbf{X}^{k+1,d}\left(\mathbf{X}^{k+1,d}\right)^T$$

Let's call

$$M_1 = \mathbf{X}^{1,k}\left(\mathbf{X}^{1,k}\right)^T \quad M_2 = \mathbf{X}^{k+1,d}\left(\mathbf{X}^{k+1,d}\right)^T$$

Thus V becomes

$$V = \mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^T\right)^{-1} M_1 \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right) + \epsilon\mathrm{Tr}\left(\left(\mathbf{X}\mathbf{X}^T\right)^{-1} M_2 \left(\mathbf{X}\mathbf{X}^T\right)^{-1}\right) =$$

Noticing that $\mathbf{X}\mathbf{X}^T = M_1 + M_2$ we get

$$= \mathrm{Tr}\left(\left(M_1 + M_2\right)^{-1} M_1 \left(M_1 + M_2\right)^{-1}\right) + \epsilon\mathrm{Tr}\left(\left(M_1 + M_2\right)^{-1} M_2 \left(M_1 + M_2\right)^{-1}\right)$$

We shall notice that $M_1, M_2 \in \mathbb{R}^{n\times n}$, they are both symmetrical and both positive semidefinite because

$$v^T AA^T v = v^T A(v^T A)^T = w^T w \geq 0$$

M_2 is a matrix defined at the product of a maximum rank matrix times its transpose, thus it's invertible.
For the first term of the sum that defines V we can apply hint 2 and we get

$$\mathrm{Tr}\left((M_1+M_2)^{-1} M_1 (M_1+M_2)^{-1}\right) \leq \mathrm{Tr}\left(M_1^{\dagger}\right) = \mathrm{Tr}\left(\left(\mathbf{X}^{1,k}\left(\mathbf{X}^{1,k}\right)^T\right)^{\dagger}\right) =$$

$$= \mathrm{Tr}\left(\left(\left(\mathbf{X}^{1,k}\right)^T\right)^{\dagger}\left(\mathbf{X}^{1,k}\right)^{\dagger}\right) = \mathrm{Tr}\left(\left(\left(\mathbf{X}^{1,k}\right)^{\dagger}\right)^T\left(\mathbf{X}^{1,k}\right)^{\dagger}\right) =$$

$$= \mathrm{Tr}\left(\left(\left(\mathbf{X}^{1,k^T}\mathbf{X}^{1,k}\right)^{-1}\mathbf{X}^{1,k^T}\right)^T\left(\mathbf{X}^{1,k^T}\mathbf{X}^{1,k}\right)^{-1}\mathbf{X}^{1,k^T}\right) = \mathrm{Tr}\left(\left(\left(\mathbf{X}^{1,k}\right)^T\mathbf{X}^{1,k}\right)^{-1}\right)$$

If now we compute the expected value we get

$$\mathbf{E}\left[\mathrm{Tr}\left(\left(\left(\mathbf{X}^{1,k}\right)^T\mathbf{X}^{1,k}\right)^{-1}\right)\right] = \mathrm{Tr}\left(\mathbf{E}\left[\left(\left(\mathbf{X}^{1,k}\right)^T\mathbf{X}^{1,k}\right)^{-1}\right]\right) = \mathrm{Tr}\left(\mathbf{E}\left[\left(\sum_{i=1}^{n} U_i U_i^T\right)^{-1}\right]\right) =$$

$$= \mathrm{Tr}\left(\frac{1}{n-k-1} I_k\right) = \frac{k}{n-k-1}$$

If we fix $0 < a < 1$ for what concerns the second term we could rewrite it as

$$\epsilon \mathrm{Tr}\left((M_1 + aM_2 + (1-a)M_2)^{-1} M_2 (M_1 + aM_2 + (1-a)M_2)^{-1}\right) =$$

$$= \frac{\epsilon}{1-a}\mathrm{Tr}\left((M_1 + aM_2 + (1-a)M_2)^{-1}(1-a)M_2 (M_1 + aM_2 + (1-a)M_2)^{-1}\right)$$

But now we can see that M_2 is positive definite because it's invertible and positive semi-definite, and M_1 is positive semi-definite, thus for $v \in \mathbb{R}^n$ we have

$$v^T (M_1 + aM_2) v = v^T M_1 v + a v^T M_2 v > 0$$

because the first term is non-negative and the second one is positive. This means that $M_1 + aM_2$ is positive definite, thus invertible.
Hence we can apply the hint and get

$$\frac{\epsilon}{1-a}\mathrm{Tr}\left((M_1 + aM_2 + (1-a)M_2)^{-1}(1-a)M_2 (M_1 + aM_2 + (1-a)M_2)^{-1}\right) =$$

$$= \frac{\epsilon}{1-a}\mathrm{Tr}\left(((1-a)M_2)^{\dagger}\right) = \frac{\epsilon}{1-a}\mathrm{Tr}\left(\frac{1}{1-a}M_2^{-1}\right) = \frac{\epsilon}{(1-a)^2}\mathrm{Tr}\left(M_2^{-1}\right) =$$

because the matrix is invertible, if now we put in the definition of M_2 we have

$$= \frac{\epsilon}{(1-a)^2}\mathrm{Tr}\left(\left(\mathbf{X}^{k+1,d}\left(\mathbf{X}^{k+1,d}\right)^T\right)^{-1}\right) = \frac{\epsilon}{(1-a)^2}\mathrm{Tr}\left(\left(\sum_{i=k+1}^{d} U_i U_i^T\right)^{-1}\right)$$

where $U_i \sim N(0, \epsilon I_n)$, if we rescale them defining $\tilde{U}_i = (1/\sqrt{\epsilon})U_i \sim N(0, I_n)$ and we take the expected value, we apply the hint

$$\mathbf{E}\left[\frac{\epsilon}{(1-a)^2}\mathrm{Tr}\left(\left(\epsilon\sum_{i=k+1}^{d}\tilde{U}_i\tilde{U}_i^T\right)^{-1}\right)\right] = \frac{1}{(1-a)^2}\frac{n}{d-k-n-1} \xrightarrow{a\to 0} \frac{n}{d-k-n-1}$$

If we put together the two upper bounds we get the result.
Exactly like in subproblem 1.4

$$\mathbf{E}_{\mathbf{X},\mathbf{z}} R\left(\hat{\omega}\right) - R\left(\omega^*\right) = \mathbf{E}_{\mathbf{X}}\left[B^2\right] + \sigma^2\mathbf{E}_{\mathbf{X}}\left[V\right]$$

The $\mathbf{E}_{\mathbf{X}}[V]$ can be bounded by 0 because

$$\frac{k}{n-k-1} \approx 0 \text{ when } n \gg k$$

$$\frac{d}{d-k-n-1} \approx 0 \text{ when } d-k \gg n$$

For what concerns the bias term, we can use the remark of subproblem 1.3.
$\mathbf{X}$ is made of Gaussian variables so they are all subgaussian.
We need only to investigate the effective rank of Σ

$$\frac{\mathrm{Tr}\left(\Sigma\right)}{\|\Sigma\|} = \frac{1}{\|\Sigma\|}\left(k + \epsilon\left(d-k\right)\right)$$

That we need $\leq d$, or rewritten

$$\|\Sigma\| \geq \frac{k+\epsilon\left(d-k\right)}{d} \approx \epsilon$$

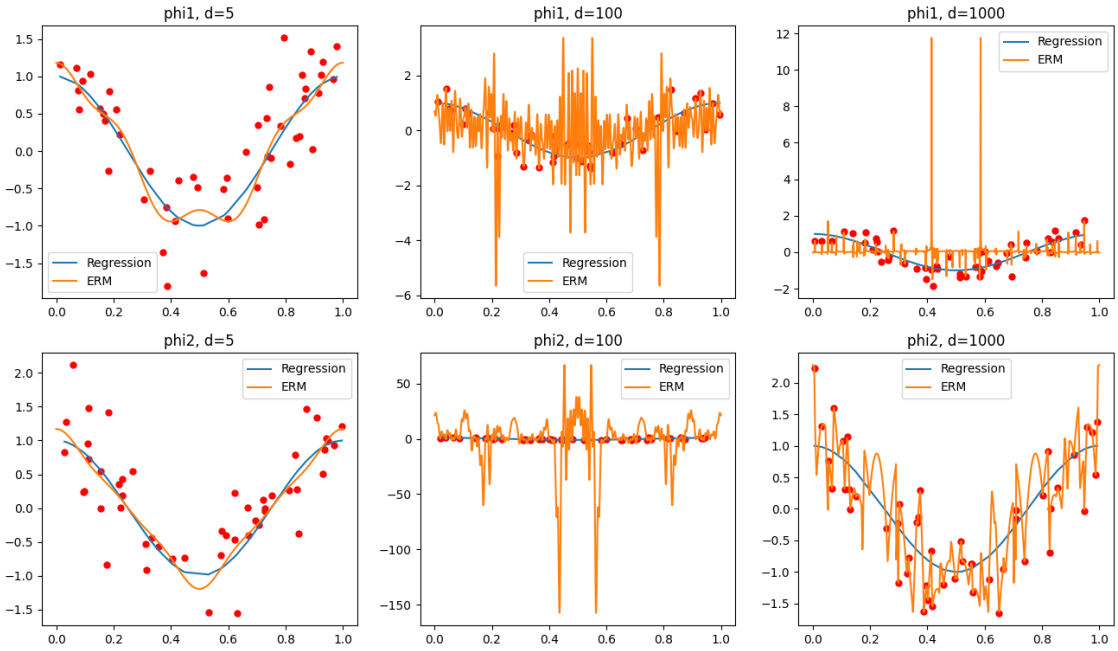
That is true because if we take the 2-norm

$$\|\Sigma\|_2^2 \leq \max\{1, \epsilon\}$$

Given these information, the remark tells us that the excess risk is vary small.

Subproblem 1.6

For what concerns $d = 5$ for both $\phi_1(x)$ and $\phi_2(x)$ we have that both the variance and the bias are low because the amount of dimensions is too small to cause overfitting and the regression is very close to the original function.
For $d = 100$, for both $\phi_1(x)$ and $\phi_2(x)$, the variance it's high but the bias it's low considering that on average the regression follows the movement of the original function.
For $d = 1000$, for $\phi_1(x)$, the bias is high because the regression is almost a constant and so not close to the function. On the other hand, for the same reason, the variance is low because, apart from some isolated points, the function is almost constant.
For $\phi_2(x)$ we see the perfect example of overfitting, where we have a low bias, because the regression is similar to the function, and a high variance, because the function has to pass through each point in the dataset, even if they are spread out.



Problem 2

Subproblem 2.1

From the condition (3) in the homework sheet, we know

$$R(\omega_{t+1}) \leq R(\omega_t) - \eta \|\nabla R(\omega_t)\|_2^2 + \frac{\beta\eta^2}{2} \|\nabla R(\omega_t)\|_2^2 \leq R(\omega_t) - \eta \|\nabla R(\omega_t)\|_2^2 + \frac{\eta}{2} \|\nabla R(\omega_t)\|_2^2 \leq$$

Where we used the inequality $\eta \leq 1/\beta$.
Using now the convexity of R, we get

$$\leq R(\omega^*) + \nabla R(\omega_t)^T (\omega_t - \omega^*) - \frac{\eta}{2} \|\nabla R(\omega_t)\|_2^2$$

Moving $R(\omega^*)$ on the LHS gives us

$$R(\omega_{t+1}) - R(\omega^*) \leq \nabla R(\omega_t)^T (\omega_t - \omega^*) - \frac{\eta}{2} \|\nabla R(\omega_t)\|_2^2 =$$

$$= \frac{1}{2\eta} \left(\|\omega_t - \omega^*\|_2^2 - \|\omega_t - \omega^* - \eta \nabla R(\omega_t)\|_2^2 \right) =$$

$$= \frac{1}{2\eta} \left(\|\omega_t - \omega^*\|_2^2 - \|\omega_{t+1} - \omega^*\|_2^2 \right)$$

Noticing that $R(\omega_{t+1}) \leq R(\omega_t)$, we can sum both sides of the inequality and get

$$\sum_{i=0}^{T-1} R(\omega_T) - R(\omega^*) \leq \sum_{i=0}^{T-1} R(\omega_{t+1}) - R(\omega^*) \leq \sum_{i=0}^{T-1} \frac{1}{2\eta} \left(\|\omega_t - \omega^*\|_2^2 - \|\omega_{t+1} - \omega^*\|_2^2 \right)$$

That gives

$$R\left(\omega_T\right) - R\left(\omega^*\right) \leq \frac{1}{2\eta T}\left\|\omega_0 - \omega^*\right\|_2^2 \leq \frac{\left\|\omega^*\right\|_2^2}{2\eta T}$$

Subproblem 2.2

If we fix

$$R(\omega) = \|\mathbf{X}\omega - \mathbf{y}\|_2^2$$

and we set ω^* as the minimizer of $R(\omega)$ (it exists because the function is continuous and convex), we get (from subproblem 2.1) that for all $T > 0$

$$R\left(\omega_T\right) - R\left(\omega^*\right) \leq \frac{\left\|\omega^*\right\|_2^2}{2\eta T}$$

We want to make this inequality as strict as possible, so we want to minimize the upper bound respect to η, to do that, given that $\eta \in (0, 1/\beta]$, we fix $\eta = 1/\beta$.
The problem is now to find the lowest β that makes the next statement true

$$R\left(\omega - \eta\nabla R\left(\omega\right)\right) \leq R\left(\omega\right) - \eta\left(1 - \frac{\eta\beta}{2}\right)\left\|\nabla R\left(\omega\right)\right\|_2^2$$

To do that, we start computing the gradient of $R(\omega)$

$$\nabla R\left(\omega\right) = \frac{1}{2n}\nabla\left[\left(\mathbf{X}\omega - \mathbf{y}\right)^T\left(\mathbf{X}\omega - \mathbf{y}\right)\right] = \frac{1}{2n}\left[2\mathbf{X}^T\mathbf{X}\omega - 2\mathbf{X}^T\mathbf{y}\right] = \frac{1}{n}\mathbf{X}^T\left[\mathbf{X}\omega - \mathbf{y}\right]$$

Now let's expand $R(\omega - \eta\nabla R(\omega))$

$$R\left(\omega - \eta\nabla R\left(\omega\right)\right) = \frac{1}{2n}\left\|\mathbf{X}\omega - \frac{1}{n}\eta\mathbf{X}^T\mathbf{X}\left[\mathbf{X}\omega - \mathbf{y}\right] - \mathbf{y}\right\|_2^2 \leq$$

$$\leq \frac{1}{2n}\left[\left\|\mathbf{X}\omega - \mathbf{y}\right\|_2^2 + \left\|\frac{1}{n}\eta\mathbf{X}^T\mathbf{X}\left[\mathbf{X}\omega - \mathbf{y}\right]\right\|_2^2\right] =$$

$$= R\left(\omega\right) + \frac{1}{2n^3}\eta^2\left[\mathbf{X}\omega - \mathbf{y}\right]^T\mathbf{X}^T\mathbf{X}\mathbf{X}\mathbf{X}^T\left[\mathbf{X}\omega - \mathbf{y}\right] = R\left(\omega\right) + \frac{1}{2n^3}\eta^2\left\|\mathbf{X}^T\left[\mathbf{X}\omega - \mathbf{y}\right]\right\|_{\mathbf{X}^T\mathbf{X}}^2 \leq$$

$$\leq R\left(\omega\right) + \frac{1}{2n^3}\eta^2\left\|\mathbf{X}^T\mathbf{X}\right\|\left\|\mathbf{X}^T\left[\mathbf{X}\omega - \mathbf{y}\right]\right\|_2^2 = R\left(\omega\right) + \frac{1}{2n}\eta^2\left\|\mathbf{X}^T\mathbf{X}\right\|\left\|\nabla R\left(\omega\right)\right\|_2^2$$

Thus now we impose

$$\frac{1}{2n}\eta^2\left\|\mathbf{X}^T\mathbf{X}\right\| \leq \eta\left(1 - \frac{\eta\beta}{2}\right) \Longrightarrow \frac{1}{2n}\eta\left\|\mathbf{X}^T\mathbf{X}\right\| \leq \left(1 - \frac{\eta\beta}{2}\right) \Longrightarrow \frac{2}{\eta} - \frac{\left\|\mathbf{X}^T\mathbf{X}\right\|}{n} \geq \beta$$

Setting now $\eta = 1/\beta$, we get

$$\beta \geq \frac{\left\|\mathbf{X}^T\mathbf{X}\right\|}{n}$$

Thus taking lowest β (or the highest η) gives

$$\eta = \frac{n}{\left\|\mathbf{X}^T\mathbf{X}\right\|}$$

Subproblem 2.3

(a) By Taylor's theorem we have that for a function f twice differentiable on the segment $[x, x+h]$ we have

$$f(x+h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^t H f(\xi) h$$

with $\xi \in [x, x+h]$.
In our case the function $R(\omega)$ is twice differentiable on the segment $[\omega, \omega - \eta \nabla R(\omega)]$ so we can write

$$R\left(\omega - \eta \nabla R\left(\omega\right)\right) = R\left(\omega\right) - \eta \nabla R\left(\omega\right)^T \nabla R\left(\omega\right) + \frac{1}{2}\eta^2 \nabla R\left(\omega\right)^T HR\left(\xi\right) \nabla R\left(\omega\right) =$$

$$= R\left(\omega\right) - \eta \left\|\nabla R\left(\omega\right)\right\|_2^2 + \frac{1}{2}\eta^2 \nabla R\left(\omega\right)^T HR\left(\xi\right) \nabla R\left(\omega\right) =$$

$$= R\left(\omega\right) - \eta \left\|\nabla R\left(\omega\right)\right\|_2^2 + \frac{1}{2}\eta^2 \left\|\nabla R\left(\omega\right)\right\|_{HR(\xi)}^2 \leq$$

$$\leq R\left(\omega\right) - \eta \left\|\nabla R\left(\omega\right)\right\|_2^2 + \frac{1}{2}\eta^2 \left\|HR\left(\xi\right)\right\| \left\|\nabla R\left(\omega\right)\right\|_2^2$$

for $\xi \in [\omega, \omega - \eta \nabla R(\omega)]$ and $\|\cdot\|$ the operator norm.
Let's compute the Hessian of $R(\omega)$ evaluated in ξ

$$\partial_{\omega_k} R\left(\xi\right) = \frac{1}{n}\sum_{i=1}^{n} \partial_{\omega_k} l\left(-\left\langle x_i, \xi\right\rangle y_i\right) = \frac{1}{n}\sum_{i=1}^{n} -y_i x_{ik} l'\left(-\left\langle x_i, \xi\right\rangle y_i\right)$$

$$\partial_{\omega_k \omega_h} R\left(\xi\right) = \partial_{\omega_h}\left(\frac{1}{n}\sum_{i=1}^{n} -y_i x_{ik} l'\left(-\left\langle x_i, \xi\right\rangle y_i\right)\right) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 x_{ik} x_{ih} l''\left(-\left\langle x_i, \xi\right\rangle y_i\right) =$$

Given that $y_i \in \{-1, 1\}$

$$= \frac{1}{n}\sum_{i=1}^{n} x_{ik} x_{ih} l''\left(-\left\langle x_i, \xi\right\rangle y_i\right)$$

If we define the matrices $\mathbf{X}^i$ such that

$$\mathbf{X}^i_{hk} = x_{ik} x_{ih}$$

We have that the Hessian can be rewritten as

$$HR(\xi) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}^i l''\left(-\left\langle x_i, \xi\right\rangle y_i\right)$$

Thus the norm can be upper-bounded like this

$$\left\|HR\left(\xi\right)\right\| \leq \frac{1}{n}\sum_{i=1}^{n} l''\left(-\left\langle x_i, \xi\right\rangle y_i\right) \left\|\mathbf{X}^i\right\| \leq \frac{1}{n}\sum_{i=1}^{n} l''\left(-\left\langle x_i, \xi\right\rangle y_i\right) \left(\sum_{k=1}^{d}\sum_{h=1}^{d} x_{ik}^2 x_{ih}^2\right)^{1/2} \leq$$

$$\leq \frac{1}{n}\sum_{i=1}^{n} l''\left(-\left\langle x_i, \xi\right\rangle y_i\right) B^2 \leq \frac{1}{n} \max_{v \in [\omega, \omega - \eta \nabla R(\omega)]} \sum_{i=1}^{n} l''\left(-\left\langle x_i, v\right\rangle y_i\right) B^2$$

Putting it all together we get the desired result.

(b) We start computing $l''(r)$

$$l''(r) = \frac{d^2}{dr^2}\log(1+\exp(r)) = \frac{d}{dr}\frac{\exp(r)}{1+\exp(r)} = \frac{\exp(r)+\exp(r)^2-\exp(r)^2}{(1+\exp(r))^2} =$$

$$= \frac{\exp(r)}{(1+\exp(r))^2}$$

To prove the wanted result we use that $l \in \mathcal{C}^\infty(\mathbb{R})$ and that if $f, g \in \mathcal{C}^\infty(\mathbb{R})$ and $\exists x_0 \in \mathbb{R}$ such that

$$g(x_0) \leq f(x_0)$$

$$g'(x) \leq f'(x)\ \forall x \in \mathbb{R}$$

we have that $g(x) \leq f(x)$ for all $x \in \mathbb{R}$.
In fact for $r = 0$ we have

$$\frac{\exp(0)}{(1+\exp(0))^2} = \frac{1}{4} < \log(2) = \log(1+\exp(0))$$

and that

$$l'''(r) = \frac{\exp(r)-\exp(r)^2}{(1+\exp(r))^3} = \frac{1}{(1+\exp(r))^2}\frac{\exp(r)}{1+\exp(r)} - \frac{\exp(r)^2}{(1+\exp(r))^3} \leq$$

$$\leq \frac{\exp(r)}{1+\exp(r)} = l'(r)$$

So we have that

$$l''(r) \leq l(r)\ \forall r \in \mathbb{R}$$

Hence we get

$$\max_{v\in[\omega,\omega-\eta\nabla R(\omega)]} \frac{1}{n}\sum_{i=1}^{n} l''\left(-\langle v, x_i\rangle y_i\right) \leq \max_{v\in[\omega,\omega-\eta\nabla R(\omega)]} \frac{1}{n}\sum_{i=1}^{n} l\left(-\langle v, x_i\rangle y_i\right)$$

Now noticing that $l''(r) > 0$ and $l'(r) > 0$ we get that $l(r)$ is convex and monotone increasing.
To this notion, we can add that the function

$$v \longmapsto -\langle v, x_i\rangle y_i$$

is affine, thus convex.
Hence we have that $l(-\langle v, x_i\rangle y_i)$ is convex and

$$\sum_{i=1}^{n} l(-\langle v, x_i\rangle y_i)$$

is convex because it's a non-negative weighted sum of convex functions.
We can conclude knowing that convex functions on limited sets obtain the maximum on the boundary of the set, thus

$$\max_{v\in[\omega,\omega-\eta\nabla R(\omega)]} \frac{1}{n}\sum_{i=1}^{n} l\left(-\langle v, x_i\rangle y_i\right) = \max_{v\in\{\omega,\omega-\eta\nabla R(\omega)\}} \frac{1}{n}\sum_{i=1}^{n} l\left(-\langle v, x_i\rangle y_i\right) =$$

$$= \max\left\{R(\omega), R\left(\omega-\eta\nabla R(\omega)\right)\right\}$$

(c) To prove $l'(r) \leq l(r)$ we use the same line of reasoning that we used before: we prove that there is a point where the functions follow the inequality and then we prove that the inequality is valid also for the derivatives.

$$l'(0) = \frac{1}{2} \leq \log(2) = l(0)$$

$$l''(r) = \frac{\exp(r)}{\left(1+\exp(r)\right)^2} = \frac{1}{\left(1+\exp(r)\right)}\frac{\exp(r)}{\left(1+\exp(r)\right)} \leq \frac{\exp(r)}{\left(1+\exp(r)\right)} = l'(r)$$

To prove that $\|R(\omega)\|_2^2 \leq B^2 R(\omega)^2$ we can expand the definition of the norm

$$\|R(\omega)\|_2^2 = \sum_{i=1}^{d}\left[\frac{1}{n}\sum_{i=1}^{n} -y_i l'\left(-\langle\omega, x_i\rangle y_i\right)\right]^2 x_i^2 = \left[\frac{1}{n}\sum_{i=1}^{n} -y_i l'\left(-\langle\omega, x_i\rangle y_i\right)\right]^2 \|x_i\|_2^2 \leq$$

$$\leq B^2\frac{1}{n^2}\left[\sum_{i=1}^{n} l'\left(-\langle\omega, x_i\rangle y_i\right)\right]^2 \leq B^2\frac{1}{n^2}\left[\sum_{i=1}^{n} l\left(-\langle\omega, x_i\rangle y_i\right)\right]^2 \leq B^2 R(\omega)^2$$

If now we consider that case $\eta \leq 1/B^2$ and $R(\omega) \leq 1$ and we start stating that $R(\omega - \eta\nabla R(\omega)) > R(\omega)$, we get

$$\frac{R\left(\omega-\eta\nabla R(\omega)\right)}{R(\omega)} \leq$$

$$\leq \frac{1}{R(\omega)}\left(R(\omega) - \eta\left\|\nabla R(\omega)\right\|_2^2 + \frac{B^2\eta^2}{2}\left\|\nabla R(\omega)\right\|_2^2 \max\left\{R(\omega), R\left(\omega-\eta\nabla R(\omega)\right)\right\}\right) \leq$$

Using the statement $R(\omega - \eta\nabla R(\omega)) \geq R(\omega)$

$$\leq 1 - \eta\frac{\|\nabla R(\omega)\|_2^2}{R\left(\omega\right)}\left(1 - \frac{B^2\eta}{2}R\left(\omega-\eta\nabla R(\omega)\right)\right)$$

If we prove that the term inside the parenthesis is ≥ 0 we get that

$$\frac{R\left(\omega-\eta\nabla R(\omega)\right)}{R(\omega)} \leq 1 - \eta\frac{\|\nabla R(\omega)\|_2^2}{R\left(\omega\right)} \leq 1$$

That is a contradiction.
The request of that term being ≥ 0 is the same as asking

$$\frac{B^2\eta}{2}R\left(\omega-\eta\nabla R(\omega)\right) \leq 1$$

Knowing that $\frac{B^2\eta}{2} \leq \frac{1}{2}$, we get that the request becomes

$$R\left(\omega - \eta\nabla R(\omega)\right) \leq 2$$

To prove it we start again from the first inequality

$$\frac{R\left(\omega - \eta\nabla R(\omega)\right)}{R(\omega)} \leq$$

$$\leq \frac{1}{R(\omega)}\left(R(\omega) - \eta\left\|\nabla R(\omega)\right\|_2^2 + \frac{B^2\eta^2}{2}\left\|\nabla R(\omega)\right\|_2^2 \max\left\{R(\omega), R\left(\omega - \eta\nabla R(\omega)\right)\right\}\right)$$

If we move the terms with the fraction $R\left(\omega - \eta\nabla R(\omega)\right)/R(\omega)$ on the left

$$\frac{R\left(\omega - \eta\nabla R(\omega)\right)}{R(\omega)}\left(1 - \frac{B^2\eta^2}{2}\left\|\nabla R(\omega)\right\|_2^2\right) \leq 1 - \eta\frac{\left\|\nabla R(\omega)\right\|_2^2}{R(\omega)}$$

Using now that

$$\frac{B^2\eta^2}{2}\left\|\nabla R(\omega)\right\|_2^2 \leq \frac{\eta}{2}B^2R(\omega)^2 \leq \frac{1}{2}$$

The LHS can be lower-bounded and we get

$$\frac{R\left(\omega - \eta\nabla R(\omega)\right)}{R(\omega)} \leq 2 - 2\eta\frac{\left\|\nabla R(\omega)\right\|_2^2}{R(\omega)} \Longrightarrow$$

$$\Longrightarrow R\left(\omega - \eta\nabla R(\omega)\right) \leq 2R(\omega) - 2\eta\left\|\nabla R(\omega)\right\|_2^2 \leq 2 - 2\eta\left\|\nabla R(\omega)\right\|_2^2 \leq 2$$

(d) From the inequality proved in (a) with the condition $\beta = B^2$ we have

$$R\left(\omega - \eta\nabla R\left(\omega\right)\right) \leq R\left(\omega\right) - \eta\left\|\nabla R\left(\omega\right)\right\|_2^2 + \frac{B^2}{2}\eta^2\left\|\nabla R\left(\omega\right)\right\|_2^2 \max\left\{R(\omega), R\left(\omega - \eta\nabla R(\omega)\right)\right\}$$

If now we prove that $R(\omega) \leq 1$ we can apply the result from (c) and get that $\max\left\{R(\omega), R\left(\omega - \eta\nabla R(\omega)\right)\right\} \leq 1$ that leads to

$$R\left(\omega - \eta\nabla R\left(\omega\right)\right) \leq R\left(\omega\right) - \eta\left\|\nabla R\left(\omega\right)\right\|_2^2 + \frac{B^2}{2}\eta^2\left\|\nabla R\left(\omega\right)\right\|_2^2 =$$

$$= R\left(\omega\right) - \eta\left(1 - \frac{\eta\beta^2}{2}\right)\left\|\nabla R\left(\omega\right)\right\|_2^2$$

Now to prove that $R(\omega) \leq 1$ we use an induction argument

$$R(0) = \frac{1}{n}\sum_{i=1}^{n} l\left(-\langle x_{,}0\rangle y_i\right) = \log(2) < 1$$

If $R(\omega_t) \leq 1$ we can use point (c) and conclude that $R(\omega_{t+1}) \leq R(\omega_t) \leq 1$.

Subproblem 2.4

(a) The function

$$f(\omega) = R(\omega_t) + \langle \nabla R(\omega_t), \omega - \omega_t \rangle + \frac{1}{2\eta} \|\omega - \omega_t\|_2^2$$

is differentiable and convex in ω (because sum of an affine and a non-negative weighted convex function), thus it has a minimizer, and it's reached when the gradient is null.

$$\nabla f(\omega) = \nabla R(\omega_t) + \frac{1}{\eta}\omega - \frac{1}{\eta}\omega_t = 0 \Longrightarrow \omega = \omega_t - \eta \nabla R(\omega_t)$$

This proves the result.

(b) If we have a function $f(\omega)$ that has $\tilde{\omega}$ as a minimizer we know that $\forall \omega \in \mathbb{R}^d$

$$f(\omega) \geq f(\tilde{\omega}) = f(\tilde{\omega}) + \mathbf{0}^T (\omega - \tilde{\omega})$$

So $\mathbf{0}$ is a subgradient.
For our specific problem, we get that if $\omega_{t+1} = \text{argmin}_{\omega \in \mathbb{R}^d}(R(\omega) + (1/2\eta)\|\omega - \omega_t\|_2^2)$

$$\mathbf{0} \in \partial R(\omega_{t+1}) + \frac{1}{\eta}(\omega_{t+1} - \omega_t) \Longrightarrow -\frac{1}{\eta}(\omega_{t+1} - \omega_t) \in \partial R(\omega_{t+1})$$

R is convex so for any $\omega^* \in \mathbb{R}^d$

$$R(\omega^*) \geq R(\omega_{t+1}) - \frac{1}{\eta}(\omega_{t+1} - \omega_t)^T (\omega^* - \omega_{t+1})$$

$$R(\omega_{t+1}) \leq R(\omega^*) + \frac{1}{\eta}(\omega_{t+1} - \omega_t)^T (\omega^* - \omega_{t+1}) = R(\omega^*) - \frac{1}{\eta}(\omega_t - \omega_{t+1})^T (\omega^* - \omega_{t+1})$$

So we have

$$R(\omega_{t+1}) - R(\omega^*) \leq -\frac{1}{\eta}(\omega_t - \omega_{t+1})^T (\omega^* - \omega_{t+1}) \leq$$

$$\leq -\frac{1}{\eta}(\omega_t - \omega_{t+1})^T (\omega^* - \omega_{t+1}) + \frac{1}{2\eta}\|\omega_t - \omega_{t+1}\|_2^2 =$$

$$= \frac{1}{2\eta}\left(\|\omega_t - \omega_{t+1} - (\omega^* - \omega_{t+1})\|_2^2 - \|\omega^* - \omega_{t+1}\|_2^2\right) =$$

$$= \frac{1}{2\eta}\left(\|\omega_t - \omega^*\|_2^2 - \|\omega^* - \omega_{t+1}\|_2^2\right)$$

Now we conclude like in subproblem 2.1 (for $T > 0$ we have $R(\omega_T) < R(\omega_{T-1})$ and we compute the telescopic sum on the RHS)

$$T(R(\omega_T) - R(\omega^*)) = \sum_{t=1}^{T} R(\omega_T) - R(\omega^*) \leq \sum_{t=1}^{T} R(\omega_t) - R(\omega^*) \leq$$

$$\leq \frac{1}{2\eta}\|\omega_0 - \omega^*\|_2^2 - \|\omega^* - \omega_T\|_2^2 \leq \frac{1}{2\eta}\|\omega^*\|_2^2$$

Thus

$$R(\omega_T) - R(\omega^*) \leq \frac{\|\omega^*\|_2^2}{2\eta T}$$

(c) If ω_{t+1} is the minimizer of

$$\Psi(\omega) + \frac{1}{2\eta} \|\omega - u_{t+1}\|_2^2$$

we have that

$$0 \in \partial\Psi(\omega_{t+1}) + \frac{1}{\eta}(\omega_{t+1} - u_{t+1}) = \partial\Psi(\omega_{t+1}) + \frac{1}{\eta}(\omega_{t+1} - \omega_t + \eta\nabla R(\omega_t)) =$$

$$= \partial\Psi(\omega_{t+1}) + \frac{1}{\eta}(\omega_{t+1} - \omega_t) + \nabla R(\omega_t)$$

If we define

$$\omega_{t+1} = \omega_t - \eta R_\eta(\omega_t)$$

we get in the previous subgradient condition

$$0 \in \partial\Psi(\omega_t - \eta R_\eta(\omega_t)) - R_\eta(\omega_t) + \nabla R(\omega_t) \Longrightarrow R_\eta(\omega_t) - \nabla R(\omega_t) \in \partial\Psi(\omega_t - \eta R_\eta(\omega_t))$$

Now we apply the β-smoothness to $R(\omega_{t+1})$

$$R(\omega_t - \eta R_\eta(\omega_t)) \leq R(\omega_t) - \eta\nabla R(\omega_t)^T R_\eta(\omega_t) + \frac{\beta\eta^2}{2}\|R_\eta(\omega_t)\|_2^2 \leq$$

using $\eta \leq 1/\beta$

$$\leq R(\omega_t) - \eta\nabla R(\omega_t)^T R_\eta(\omega_t) + \frac{\eta}{2}\|R_\eta(\omega_t)\|_2^2$$

Now we study an upper-bound for the function $\tilde{R}(\omega)$

$$\tilde{R}(\omega_{t+1}) \leq R(\omega_t) - \eta\nabla R(\omega_t)^T R_\eta(\omega_t) + \frac{\eta}{2}\|R_\eta(\omega_t)\|_2^2 + \Psi(\omega_t - \eta R_\eta(\omega_t)) \leq$$

Using the convexity of R and of Ψ and the subgradient that we found before, we have

$$\leq R(\omega^*) + \nabla R(\omega_t)^T(\omega_t - \omega^*) - \eta\nabla R(\omega_t)^T R_\eta(\omega_t) + \frac{\eta}{2}\|R_\eta(\omega_t)\|_2^2 +$$

$$+\Psi(\omega^*) + (R_\eta(\omega_t) - \nabla R(\omega_t))^T(\omega_t - \omega^* - \eta R_\eta(\omega_t)) =$$

$$= \tilde{R}(\omega^*) + R_\eta(\omega_t)^T(\omega_t - \omega^*) - \frac{\eta}{2}\|R_\eta(\omega_t)\|_2^2$$

This inequality becomes

$$\tilde{R}(\omega_{t+1}) - \tilde{R}(\omega^*) \leq R_\eta(\omega_t)^T(\omega_t - \omega^*) - \frac{\eta}{2}\|R_\eta(\omega_t)\|_2^2 =$$

$$= \frac{1}{2\eta}\left(\|\omega_t - \omega^*\|_2^2 - \|\omega_t - \omega^* - \eta R_\eta(\omega_t)\|_2^2\right) = \frac{1}{2\eta}\left(\|\omega_t - \omega^*\|_2^2 - \|\omega_{t+1} - \omega^*\|_2^2\right)$$

And we conclude as in (b) using the telescopic sum.

(d) The problem that we are solving can be rewritten as

$$\omega_{t+1} = \underset{\omega\in\mathbb{R}^d}{\operatorname{argmin}} \left\{\lambda\|\omega\|_1 + \frac{1}{2\eta}\|\omega - u_{t+1}\|_2^2\right\}$$

We know that there is a minimizer because the function is convex and continuous, so let's study the properties that it has to follow. A necessary condition to be a minimizer of a convex function is that

$$0 \in \lambda \partial \left\| \omega_{t+1} \right\|_1 + \frac{1}{\eta} \omega_{t+1} - \frac{1}{\eta} u_{t+1}$$

So if $(\omega_{t+1})_i \neq 0$ the subgradient is the gradient itself, so

$$0 = \lambda \operatorname{sgn}\left((\omega_{t+1})_i\right) + \frac{1}{\eta}\left(\omega_{t+1}\right)_i - \frac{1}{\eta}\left(u_{t+1}\right)_i \Longrightarrow \left(\omega_{t+1}\right)_i = \left(u_{t+1}\right)_i - \lambda\eta \operatorname{sgn}\left((\omega_{t+1})_i\right)$$

We can notice that

$$\left(\omega_{t+1}\right)_i > 0 \Longrightarrow \left(u_{t+1}\right)_i = \left(\omega_{t+1}\right)_i + \lambda\eta \Longrightarrow \left(u_{t+1}\right)_i > 0$$

and

$$\left(\omega_{t+1}\right)_i < 0 \Longrightarrow \left(u_{t+1}\right)_i = \left(\omega_{t+1}\right)_i - \lambda\eta \Longrightarrow \left(u_{t+1}\right)_i < 0$$

thus

$$\operatorname{sgn}\left((\omega_{t+1})_i\right) = \operatorname{sgn}\left((u_{t+1})_i\right)$$

If $\left(\omega_{t+1}\right)_i = 0$ the condition on the subgradient becomes

$$0 \in \lambda[-1,1] - \frac{1}{\eta}\left(u_{t+1}\right)_i \Longleftrightarrow \left(u_{t+1}\right)_i \in [-\eta\lambda, \eta\lambda] \Longleftrightarrow \left|\left(u_{t+1}\right)_i\right| \leq \eta\lambda$$

This is also a sufficient condition because if $\left|\left(u_{t+1}\right)_i\right| \leq \eta\lambda$, if by contradiction we set $\left(\omega_{t+1}\right)_i = 0$ from the condition on the subgradient we get

$$\left(\omega_{t+1}\right)_i + \lambda\eta \operatorname{sgn}\left((\omega_{t+1})_i\right) = \left(u_{t+1}\right)_i \Longrightarrow \left|\left(\omega_{t+1}\right)_i + \lambda\eta \operatorname{sgn}\left((\omega_{t+1})_i\right)\right| > \eta\lambda$$

We can conclude that a closed form for $\left(\omega_{t+1}\right)_i$ is

$$\left(\omega_{t+1}\right)_i = \begin{cases} 0 & \left|\left(u_{t+1}\right)_i\right| \leq \eta\lambda \\ \left(u_{t+1}\right)_i - \lambda\eta \operatorname{sgn}\left((u_{t+1})_i\right) & \text{otherwise} \end{cases}$$

Where $\left(u_{t+1}\right)_i$ can be computed as follows

$$\nabla R\left(\omega\right) = \frac{1}{n}\left[\mathbf{X}^T\mathbf{X}\omega - \mathbf{X}^T\mathbf{y}\right]$$

$$\left(u_{t+1}\right)_i = \left(\omega_t\right)_i - \frac{\eta}{n}\left(\mathbf{X}^T\mathbf{X}\omega_t\right)_i + \frac{\eta}{n}\left(\mathbf{X}^T\mathbf{y}\right)_i$$