Queen Mary
University of London

MTH786, Semester A, 2023/24
Solutions of coursework 4 **N. Perra**

Problem 1. For this exercise we consider ridge regression problems of the form

$$\mathbf{w}_\alpha = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2 + \frac{\alpha}{2}\|\mathbf{w}\|^2 \right\}, \tag{1}$$

for data $\mathbf{y}\in\mathbb{R}^s$, a data matrix $\mathbf{X}\in\mathbb{R}^{s\times(d+1)}$ and a regularisation parameter $\alpha>0$.

1. Compute the ridge regression solution for the original data samples given in the previous question on unstable regression problems. I.e., consider data points $(x^{(1)},y^{(1)})$ with $x^{(1)}=-c$ and $y^{(1)}=2$, $(x^{(2)},y^{(2)})$ with $x^{(2)}=0$ and $y^{(2)}=2$, and $(x^{(3)},y^{(3)})$ with $x^{(3)}=c$ and $y^{(3)}=2$, for some constant $c>0$.

2. Consider validation data of the form $\left(x^{(4)},y^{(4)}\right)$, where $x^{(4)}=2$, $y^{(4)}=1$. Choose the regularisation parameter α such that it solves the minimisation problem for validation error, i.e.

$$\hat{\alpha} = \arg\min_{\alpha\geq 0} \left\{ \left| w_\alpha^{(0)} + w_\alpha^{(1)} x^{(4)} - y^{(4)} \right|^2 \right\}.$$

3. Repeat the same exercise for the perturbed data samples, i.e. $\mathbf{y}_\delta$ that reads $y_\delta^{(1)}=2+\varepsilon$, $y_\delta^{(2)}=2+\varepsilon$ and $y_\delta^{(3)}=2-\varepsilon$.

Solutions:

1. We proceed as in the related exercise and compute the ridge regression optimality condition $(\mathbf{X}^\top\mathbf{X}+\alpha)\mathbf{w}_\alpha=\mathbf{X}^\top\mathbf{y}$, which for this example reads

$$\left(\begin{pmatrix} 3 & 0 \\ 0 & 2c^2 \end{pmatrix} + \alpha \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \mathbf{w}_\alpha = \mathbf{X}^\top\mathbf{y},$$

We easily invert the matrix on the left-hand-side to compute

$$\begin{aligned}
\mathbf{w}_\alpha &= \begin{pmatrix} \frac{1}{\alpha+3} & 0 \\ 0 & \frac{1}{2c^2+\alpha} \end{pmatrix} \mathbf{X}^\top\mathbf{y} \\
&= \begin{pmatrix} \frac{1}{\alpha+3} & \frac{1}{\alpha+3} & \frac{1}{\alpha+3} \\ -\frac{c}{2c^2+\alpha} & 0 & \frac{c}{2c^2+\alpha} \end{pmatrix} \mathbf{y} \\
&= \begin{pmatrix} \frac{1}{\alpha+3} & \frac{1}{\alpha+3} & \frac{1}{\alpha+3} \\ -\frac{c}{2c^2+\alpha} & 0 & \frac{c}{2c^2+\alpha} \end{pmatrix} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} \\
&= \begin{pmatrix} \frac{6}{\alpha+3} \\ 0 \end{pmatrix}.
\end{aligned}$$

2. Now we proceed with the calculation of a validation error. Plugging the weights we have found above one gets

$$\mathrm{L}_v = \left(\frac{6}{\alpha+3} - 1\right)^2 = \left(\frac{3-\alpha}{\alpha+6}\right)^2,$$

which is obviously minimized for $\alpha = \hat{\alpha} = 3$.

Remark: This answer should not be of a big surprise to you. It first seems that we have got a wrong prediction function of the form $f(x) = w^{(0)}_{\alpha=3} + w^{(1)}_{\alpha=3} x = 1$. Instead of previously (in the previous assignment) obtaining $f(x) = w^{(0)}_{\alpha=0} + w^{(1)}_{\alpha=0} x = 2$. But in fact what has happened is the following:

- by first studying the ridge regression we ask the question what is the optimal curve fitting our initial three points with the penalty given for bad fitting or large weights. And the answer is the straight line parallel to x-axis, with the y-coordinate depending on the ratio between two penalties (miss fit and weights size).
- later we ask to find such a line that is the closest one to the point $(2,1)$. Which is obviously $f(x) = 1$ line.

3. In identical fashion to the previous exercise we compute

$$\mathbf{w}_\alpha = \begin{pmatrix} \frac{1}{\alpha+3} & \frac{1}{\alpha+3} & \frac{1}{\alpha+3} \\ -\frac{c}{2c^2+\alpha} & 0 & \frac{c}{2c^2+\alpha} \end{pmatrix} \begin{pmatrix} 2+\varepsilon \\ 2+\varepsilon \\ 2-\varepsilon \end{pmatrix} = \begin{pmatrix} \frac{\varepsilon+6}{\alpha+3} \\ -\frac{2c\varepsilon}{2c^2+\alpha} \end{pmatrix}$$

For the validation error one has

$$\begin{aligned} \mathrm{L}_v &= \left(\frac{6+\varepsilon}{\alpha+3} - \frac{2c\varepsilon}{2c^2+\alpha}\cdot 2 - 1\right)^2 = \left(\frac{3+\varepsilon-\alpha}{\alpha+6} - \frac{4c\varepsilon}{2c^2+\alpha}\right)^2 \\ &= \left(\frac{-\alpha^2 - (2c^2 - \varepsilon - 3 + 4c\varepsilon)\,\alpha + 6c^2 + 2c^2\varepsilon - 24c\varepsilon}{(\alpha+3)(\alpha+2c^2)}\right)^2. \end{aligned}$$

Let us denote

$$P(\alpha) = \alpha^2 + \left(2c^2 - \varepsilon - 3 + 4c\varepsilon\right)\alpha - 2c\left(3c + c\varepsilon - 12\varepsilon\right), \quad Q(\alpha) = \alpha^2 + \alpha\left(2c^2+3\right)\alpha + 6c^2.$$

Obviously, if $P(\alpha)$ can take a value 0, then the minimum validation error would be zero. To understand whether the polynomial P is taking zero value one needs to analyse the quadratic polynomial. Below we consider ε being small, as this is a noise term. If $c \gg \varepsilon$, then

$$P(\alpha) \approx \alpha^2 - \alpha\left(3 + 2c^2\right) - 6c^2 \Rightarrow \hat{\alpha} \approx 3.$$

Now let us consider the case of $c \ll \varepsilon$, which is the case that leads to a huge change in weights when the linear regression studied. In this case

$$P(\alpha) \approx \alpha^2 - \alpha(3+\varepsilon) + 24c\varepsilon \Rightarrow \hat{\alpha} \approx 3 + \varepsilon - \frac{24c\varepsilon}{3+\varepsilon}.$$

In both cases, $\hat{\alpha}$ doesn't change a lot from its value 3. Thus we can see that the ridge regression is stable under the data perturbation.

Problem 2. For this exercise we consider ridge regression problems of the form

$$\mathbf{w}_\alpha = \arg\min_{\mathbf{w}\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\left\|\mathbf{X}\mathbf{w}-\mathbf{y}\right\|^2 + \frac{\alpha}{2}\left\|\mathbf{w}\right\|^2 \right\}, \tag{2}$$

for data output $\mathbf{y}\in\mathbb{R}^s$, a data matrix $\mathbf{X}\in\mathbb{R}^{s\times(d+1)}$ and a regularisation parameter $\alpha>0$.

1. Calculate the gradient of the energy function $E(\mathbf{w}) = \frac{1}{2}\left\|\mathbf{X}\mathbf{w}-\mathbf{y}\right\|^2 + \frac{\alpha}{2}\left\|[w]\right\|^2$.
2. Prove that $E(\mathbf{w})$ is a convex and bounded from below function.
3. Combine the above results to conclude that there is a unique solution $\mathbf{w}_\alpha$ of the minimisation problem (2) which also solves the normal equation
$$\left(\mathbf{X}^\top\mathbf{X}+\alpha I\right)\mathbf{w}_\alpha = \mathbf{X}^\top\mathbf{y}.$$
4. Continuously differentiable function $f:\mathbb{R}^{d+1}\to\mathbb{R}$ is called L-smooth if
$$\left\|\nabla f(\mathbf{u})-\nabla f(\mathbf{v})\right\| \le L\left\|\mathbf{u}-\mathbf{v}\right\|,$$
for any vectors $\mathbf{u},\mathbf{v}\in\mathbb{R}^{d+1}$. Prove that the energy function E is L-smooth for some value of L. Try to identify the smallest possible such a value L.

Solutions:

1. The energy function $E(\mathbf{w})$ could be rewritten as

$$\begin{aligned} E\left(w^{(0)},w^{(1)},\ldots,w^{(d)}\right) &= \frac{1}{2}\sum_{j=1}^{s}\left(\mathbf{X}\mathbf{w}-\mathbf{y}\right)_j^2 + \frac{\alpha}{2}\sum_{j=0}^{d}\left(w^{(j)}\right)^2 \\ &= \frac{1}{2}\sum_{j=1}^{s}\left(\sum_{i=0}^{d}\mathbf{X}_{j,i}w^{(i)}-y_j\right)^2 + \frac{\alpha}{2}\sum_{j=0}^{d}\left(w^{(j)}\right)^2. \end{aligned}$$

Then the partial derivative with respect to $w^{(p)}$, for some $0\le p\le d$ is equal to

$$\begin{aligned} \frac{\partial}{\partial w^{(p)}}E(\mathbf{w}) &= \frac{1}{2}\sum_{j=1}^{s}\frac{\partial}{\partial w^{(p)}}\left(\sum_{i=0}^{d}\mathbf{X}_{j,i}w^{(i)}-y_j\right)^2 + \frac{\alpha}{2}\sum_{j=0}^{d}\frac{\partial}{\partial w^{(p)}}\left(w^{(j)}\right)^2 \\ &= \sum_{j=1}^{s}\mathbf{X}_{j,p}\left(\sum_{i=0}^{d}\mathbf{X}_{j,i}w^{(i)}-y_j\right)+\alpha w^{(p)} \\ &= \sum_{j=1}^{s}\sum_{i=0}^{d}\mathbf{X}^\top_{p,j}\mathbf{X}_{j,i}w^{(i)} - \sum_{j=1}^{s}\mathbf{X}^\top_{p,j}y_j + \alpha w^{(p)} \\ &= \left(\mathbf{X}^\top\mathbf{X}\mathbf{w}-\mathbf{X}^\top\mathbf{y}+\alpha\mathbf{w}\right)_p. \end{aligned}$$

Thus the gradient is then equal

$$\nabla E(\mathbf{w}) = \left(\mathbf{X}^\top\mathbf{X}+\alpha I\right)\mathbf{w}-\mathbf{X}^\top\mathbf{y}.$$

2. We have previously shown (see previous Assignments) that:

- $MSE(\mathbf{w}) = \frac{1}{2}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ is a convex function;
- $\|\mathbf{w}\|^2$ is a strictly convex function, and thus for $\alpha > 0$ $\frac{\alpha}{2}\|\mathbf{w}\|^2$ is strictly convex;
- the sum of two convex functions is convex.

When combined all together this yields that $E(\mathbf{w})$ is strictly convex.

3. Energy function $E(\mathbf{w})$ is strictly convex and is bounded from below by $E(\mathbf{w}) \geq 0$. Function $E(\mathbf{w})$ is also continuously differentiable. Therefore (see Lecture notes),
   - there exist the unique minimizer $\mathbf{w}_\alpha = \arg\min E(\mathbf{w})$;
   - and this minimizer is the unique solution of $\nabla E(\mathbf{w}) = 0$.

   This finishes the proof.

4. To prove the energy function $E(w)$ is L-smooth one needs to evaluate the value of

$$\begin{aligned}\Delta_{\mathbf{w},\mathbf{w}'} := \nabla E(\mathbf{w}) - \nabla E(\mathbf{w}') &= \mathbf{X}^\top\mathbf{X}\mathbf{w} + \alpha\mathbf{w} - \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}\mathbf{w}' - \alpha\mathbf{w}' + \mathbf{X}^\top\mathbf{y} \\ &= \left(\mathbf{X}^\top\mathbf{X} + \alpha I\right)(\mathbf{w} - \mathbf{w}').\end{aligned}$$

   The best we can do to estimate a norm of the right hand side is to use a bound via matrix norm

$$\|\Delta_{\mathbf{w},\mathbf{w}'}\| \leq \left\|\mathbf{X}^\top\mathbf{X} + \alpha I\right\| \|\mathbf{w} - \mathbf{w}'\|.$$

   Now, defining $L = \left\|\mathbf{X}^\top\mathbf{X} + \alpha I\right\|$ we obtain a necessary inequality.

   Remark: The value of L can be also written as $L = \sigma_1^2 + \alpha$, where σ_1 is the largest singular value of matrix $\mathbf{X}$. This value of L is indeed an optimal one, because if $\mathbf{w} - \mathbf{w}'$ is parallel to a corresponding right singular vector of $\mathbf{X}$ we would indeed have

$$\Delta_{\mathbf{w},\mathbf{w}'} = L(\mathbf{w} - \mathbf{w}').$$

Problem 3. Suppose we are given s data vectors $\{\mathbf{y}^{(j)}\}_{j=1}^s$ where each $\mathbf{y}^{(j)} \in \mathbb{R}^n$ is of the form $\mathbf{y}^{(j)} = \mathbf{X}\mathbf{w}^\dagger + \varepsilon^{(j)}$, for weights $\mathbf{w}^\dagger$ and outcomes $\varepsilon^{(1)}, \ldots, \varepsilon^{(s)}$ of a random vectors ε with its expectation being zero, i.e. $\mathbb{E}\left[\varepsilon^{(\mathbf{j})}\right] = 0$.

1. Show that the expected value of a random variable is linear.
2. Show that the expected value of a constant value is that constant value itself.
3. Compute the expectation of $\hat{\mathbf{w}}^{(j)}$, where $\hat{\mathbf{w}}^{(j)}$ are the solutions of $\mathbf{X}^\top\mathbf{X}\hat{\mathbf{w}}^{(j)} = \mathbf{X}^\top\mathbf{y}^{(j)}$.
4. Compute the expectation of $\hat{\mathbf{w}}_\alpha^{(j)}$, where $\hat{\mathbf{w}}_\alpha^{(j)}$ are the solutions of (2) for data $\mathbf{y}^{(j)}$.
5. Can the two previous expectations match for any value other than $\alpha = 0$?

Solutions:

1. (a) For two random variables X and Y with outcomes $\{x_n\}_{n=1}^m$ and $\{y_n\}_{n=1}^m$ we immediately observe

$$\begin{aligned}\mathbb{E}_n[x_n + y_n] &= \sum_{n=1}^{m}(x_n + y_n)\rho_n = \left(\sum_{n=1}^{m} x_n\,\rho_n\right) + \left(\sum_{n=1}^{m} y_n\,\rho_n\right) \\ &= \mathbb{E}_n[x_n] + \mathbb{E}_n[y_n]\,,\end{aligned}$$

due to the linearity of the sum.

(b) For a constant, such as $Xw^\dagger$, we immediately observe

$$\begin{aligned}\mathbb{E}_n[Xw^\dagger] &= \sum_{n=1}^{m}(Xw^\dagger)\,\rho_n = Xw^\dagger \sum_{n=1}^{m}\rho_n \\ &= Xw^\dagger\,,\end{aligned}$$

due to $\sum_{n=1}^m \rho_n = 1$.

(c) Note that based on the two previous results we compute $\mathbb{E}_n[y_n] = \mathbb{E}_n[Xw^\dagger + \varepsilon_n] = Xw^\dagger + \mathbb{E}_n[\varepsilon_n] = 0$. For deterministic X we then obtain

$$\begin{aligned}\mathbb{E}_n[\hat{w}_n] &= \mathbb{E}_n[(X^\top X)^{-1}X^\top y_n] = (X^\top X)^{-1}X^\top \mathbb{E}_n[y_n] = (X^\top X)^{-1}X^\top X w^\dagger \\ &= w^\dagger\,.\end{aligned}$$

(d) For the ridge regression solution w_α we compute the following expectation

$$\begin{aligned}\mathbb{E}_n[(w_\alpha)_n] &= \mathbb{E}_n\left[(X^\top X + \alpha I)^{-1}X^\top y_n\right] \\ &= (X^\top X + \alpha I)^{-1}X^\top \mathbb{E}_n\left[y_n\right] \\ &= (X^\top X + \alpha I)^{-1}X^\top X w^\dagger \\ &= w^\dagger - \alpha(X^\top X + \alpha I)^{-1}w^\dagger\,.\end{aligned}$$

(e) Unless α is chosen to be $\alpha = 0$, we can never match the two expectations, since

$$\begin{aligned}\mathbb{E}_n[\hat{w}_n] - \mathbb{E}_n[(w_\alpha)_n] &= \mathbb{E}_n[\hat{w}_n - (w_\alpha)_n] \\ &= \alpha(X^\top X + \alpha I)^{-1}w^\dagger\,,\end{aligned}$$

which is only zero for arbitrary $w^\dagger$ if $\alpha = 0$.