Queen Mary
University of London

**MTH786U/P, Semester C, 2020/21**

**Assignment 4 Solutions** **M. Poplavskyi**

---

This week marks a mixture of theoretical and practical coursework. Please make sure that your Python environment is all set up.

# Ridge regression

For this exercise we consider ridge regression problems of the form

$$w_\alpha = \arg\min_{w\in\mathbb{R}^{d+1}} \left\{ \frac{1}{2}\|Xw - y\|^2 + \frac{\alpha}{2}\|w\|^2 \right\} , \qquad (1)$$

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

1. Calculate the gradient of the energy function $E(w) = \frac{1}{2}\|Xw - y\|^2 + \frac{\alpha}{2}\|w\|^2$.
2. Prove that $E(w)$ is a convex and bounded from below function.
3. Combine the above results to conclude that there is a unique solution $w_\alpha$ of the minimisation problem (1) which also solves the normal equation

$$\left(X^\top X + \alpha I\right) w_\alpha = X^\top y.$$

4. Continuously differentiable function $f : \mathbb{R}^{d+1} \to \mathbb{R}$ is called $L$-smooth if

$$\|\nabla f(u) - \nabla f(v)\| \le L\|u - v\| ,$$

for any vectors $u, v \in \mathbb{R}^{d+1}$. Prove that the energy function $E$ is $L$-smooth for some value of $L$. Try to identify the smallest possible such a value $L$.

**Solution**:

1. The energy function $E(w)$ could be rewritten as

$$\begin{aligned} E\left(w^{(0)}, w^{(1)}, \dots, w^{(d)}\right) &= \frac{1}{2}\sum_{j=1}^{s}\left(w^{(0)} + w^{(1)}x_1^{(j)} + \dots + w^{(d)}x_d^{(j)} - y^{(j)}\right)^2 \\ &\quad + \frac{\alpha}{2}\sum_{j=0}^{d}\left(w^{(j)}\right)^2 . \end{aligned}$$

Then the gradient is equal to

$$
\begin{aligned}
\nabla E(w) = \quad & \Bigg( \sum_{j=1}^{s} \left( w^{(0)} + w^{(1)} x_1^{(j)} + \ldots + w^{(d)} x_d^{(j)} - y^{(j)} \right) + \alpha w^{(0)}, \\
& \sum_{j=1}^{s} x_1^{(j)} \left( w^{(0)} + w^{(1)} x_1^{(j)} + \ldots + w^{(d)} x_d^{(j)} - y^{(j)} \right) + \alpha w^{(1)} \\
& \ldots, \\
& \sum_{j=1}^{s} x_d^{(j)} \left( w^{(0)} + w^{(1)} x_1^{(j)} + \ldots + w^{(d)} x_d^{(j)} - y^{(j)} \right) + \alpha w^{(d)} \Bigg).
\end{aligned}
$$

This can be equivalently rewritten as

$$\nabla E(w) = \left( X^\top X + \alpha I \right) w - X^\top y,$$

where

$$X = \begin{pmatrix} 1 & x_1^{(1)} & \ldots & x_d^{(1)} \\ 1 & x_1^{(2)} & \ldots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(s)} & \ldots & x_d^{(s)} \end{pmatrix}, \qquad w = \begin{pmatrix} w^{(0)} \\ w^{(1)} \\ \vdots \\ w^{(d)} \end{pmatrix}, \qquad y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(s)} \end{pmatrix}.$$

Indeed,

$$
\begin{aligned}
\left( X^\top X w - X^\top y + \alpha I w \right)_p &= \sum_{j,k} X^\top_{p,j} X_{j,k} w_k - \sum_j X^\top_{p,j} y_j + \alpha w^{(p)} \\
&= \sum_{j=1}^{s} \sum_{k=0}^{d} x_p^{(j)} x_k^{(j)} w^{(k)} - \sum_{j=1} x_p^{(j)} y^{(j)} + \alpha w^{(p)}.
\end{aligned}
$$

2. We have previously shown (see Assignment 2) that:

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

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

3. Energy function $E(w)$ is strictly convex and is bounded from below by $E(w) \geq 0$. Function $E(w)$ is also continuously differentiable. Therefore (see Lecture notes),

   - there exist the unique minimizer $w_\alpha = \arg\min E(w)$;
   - and this minimizer is the unique solution of $\nabla E(w) = 0$.

   This finishes the proof.

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

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

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

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

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

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

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