In this problem sheet, we consider the problem of linear regression with d predictors and one intercept,

$$\mathbf{y} = \beta_0 + \mathbf{X}\beta + \epsilon\,,$$

where $\mathbf{y}^t = (y_1, \ldots, y_n)$ is the column vector of target values, $\beta^t = (\beta_1, \ldots, \beta_d)$ is the column vector of coefficients excluding the intercept, $\epsilon^t = (\epsilon_1, \ldots, \epsilon_n)$ is the vector of random errors, and $\mathbf{X}$ is the $n \times d$ matrix of (standardised) observations given by

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots & x_{1d} \\ x_{21} & \ldots & x_{2d} \\ \vdots & & \vdots \\ x_{n1} & \ldots & x_{nd} \end{pmatrix}.$$

We are looking for the solution minimising the penalised sum of squares

$$RSS_p(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^t(\mathbf{y} - \mathbf{X}\beta) + \lambda||\beta||_p^p\,, \tag{1}$$

where $\lambda > 0$, and $p = 1$ (lasso), $p = 2$ (ridge regression), etc.

Problem 0.

(i) Consider the ridge regression problem,

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{d} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{d} \beta_j^2 \right\}.$$

Show that this problem is equivalent to the problem

$$\hat{\beta}^s_{\text{ridge}} = \arg\min_{\beta^s} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0^s - \sum_{j=1}^{d} (x_{ij} - \bar{x}_j)\beta_j^s \right)^2 + \lambda \sum_{j=1}^{d} (\beta_j^s)^2 \right\}.$$

Give the correspondence between β^s and the original β. Characterise the solution to this modified criterion, and explain why the intercept does not appear in the expression of the $RSS(\lambda)$ given in (1).

(ii) Show that the vector β minimising (1) is $\hat{\beta}_\lambda = (\mathbf{X}^t\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^t\mathbf{y}$. Argue why the matrix $\mathbf{X}^t\mathbf{X} + \lambda I_d$ is always positive definite, irrespectively of the rank of $\mathbf{X}$.

(iii) Suppose that $\mathbf{X}^t\mathbf{X} = I_d$. Characterise the ridge and lasso solutions in this case, and express them in terms of the least squares solution. Explain why ridge performs a proportional shrinkage, while lasso performs soft-thresholding.

(iv) Recall the dimensions and properties of the matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{\Lambda}$ in the SVD decomposition of $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t$, for a $n \times d$ matrix of rank d, with $d < n$.

(v) The solution to the ridge regression problem is $\hat{\beta}_\lambda = (\mathbf{X}^t\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^t\mathbf{y}$, where $\lambda > 0$ is the tuning parameter. Show that the ridge estimate can be written

$$\hat{\mathbf{y}}_\lambda := \mathbf{X}\hat{\beta}_\lambda = \sum_{j=1}^{d} \frac{\lambda_j^2}{\lambda_j^2 + \lambda} < \mathbf{u}_j, \mathbf{y} > \mathbf{u}_j \,,$$

where the λ_js are the diagonal entries of Λ, and $\mathbf{u}_j$ the columns of $\mathbf{U}$.

(vi) Derive the eigenvalue-eigenvector pairs of $\mathbf{X}^t\mathbf{X}$.

(vii) Assuming the columns of $\mathbf{X}$ are centered, deduce the eigenvalue-eigenvector pairs of the sample covariance matrix $\mathbf{S} = \mathbf{X}^t\mathbf{X}/n$.

(viii) Deduce a geometrical interpretation of the λ_j, and a geometrical interpretation of the ridge estimates. In which directions does ridge regression shrink the coefficients the most?

Problem 1.

(i) Recall the expression of the ridge regression solution $\hat{\beta}_\lambda$, minimising the penalised sum of squares (1).

(ii) Show that $\hat{\beta}_\lambda$ is a biased estimate of β.

(iii) The ridge fits are given by $\hat{\mathbf{y}}_\lambda = \mathbf{X}\hat{\beta}_\lambda = \mathbf{H}_\lambda\mathbf{y}$. Recall the expression of the matrix $\mathbf{H}_\lambda$. Show that $\mathbf{H}_\lambda$ is not a projection matrix for $\lambda > 0$.

Hint: Show that $\mathbf{H}_\lambda$ is not idempotent.

(iv) Show that the ridge fits $\hat{\mathbf{y}}_\lambda$ are not perpendicular to the ridge residuals $\hat{\epsilon}(\lambda) := \mathbf{y} - \hat{\mathbf{y}}_\lambda$.

Problem 2.
Show that the ridge solution can be re-expressed as a lest square solution of a modified dataset.

Problem 3.
Suppose that $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^t$ is of rank $r \leq d$ and put $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$, where $\mathbf{u}_i \in \mathbb{R}^n$ and $\mathbf{v}_i \in \mathbb{R}^d$.

(i) Put $\hat{\beta}_\lambda = (\mathbf{X}^t\mathbf{X} + \lambda I_d)^{-1}\mathbf{X}^t\mathbf{y} =: \mathbf{A}_\lambda\mathbf{y}$. Show that

$$\mathbf{A}_\lambda = \sum_{j=1}^{r} \frac{\lambda_j}{\lambda_j^2 + \lambda}\mathbf{v}_j\mathbf{u}_j^t \,.$$

(ii) Put $\mathbf{P} := \sum_{j=1}^{r} \mathbf{v}_j \mathbf{v}_j^t$, the projection matrix onto the row space of $\mathbf{X}$. Check that we have

$$\mathbf{E}\hat{\beta}_\lambda = \sum_{j=1}^{r} \frac{\lambda_j^2}{\lambda_j^2 + \lambda} \langle \mathbf{v}_j,\, \beta \rangle \mathbf{v}_j \,,$$

and

$$||\beta - \mathbf{E}\hat{\beta}_\lambda||_2^2 = \beta^t(\mathbf{I}_d - \mathbf{P})\beta + \sum_{j=1}^{r} \left(\frac{\lambda}{\lambda + \lambda_j^2} \right)^2 \langle \mathbf{v}_j,\, \beta \rangle^2 \,.$$

What is the value of $||\beta - \mathbf{E}\hat{\beta}_\lambda||_2^2$ if $r = d$ and $\lambda = 0$?

(iii) Check that the total variance $\mathrm{var}\hat{\beta}_\lambda := \sum_{j=1}^{d} \mathrm{var}(\hat{\beta}_\lambda)_j$ is given by

$$\mathrm{var}\hat{\beta}_\lambda = \sigma^2 \mathrm{Tr}(\mathbf{A}_\lambda^t \mathbf{A}_\lambda) = \sigma^2 \sum_{j=1}^{r} \left(\frac{\lambda_j}{\lambda_j^2 + \lambda} \right)^2 .$$

(iv) How do the square bias and the variance of $\hat{\beta}_\lambda$ vary as λ increases?

Problem 4.
The goal of this problem is to understand the ridge and lasso solutions from a Bayesian point of view. Consider the Gaussian sampling model $\mathbf{y} \,|\, \beta \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2\mathbf{I})$.

(i) Suppose we put a normal prior distribution on the parameters, $\beta \sim \mathcal{N}(0, \tau\mathbf{I})$, for some $\tau > 0$. Show that the posterior distribution of β given $\mathbf{y}$ is normal with mean μ and covariance matrix Σ, where

$$\Sigma^{-1} = \frac{1}{\tau} I + \frac{1}{\sigma^2} \mathbf{X}^t \mathbf{X} \,,$$
$$\mu = \frac{1}{\sigma^2} \Sigma \mathbf{X}^t \mathbf{y} \,.$$

Deduce from *(i)* that the ridge regression estimate is the mean (and mode) of the posterior distribution. Find the relationship between the regularization parameter λ in the ridge formula (1), and the variances τ and σ^2.

(ii) Consider now a Laplace prior on β, so that the density of β is given by

$$\left(\frac{\tau}{2} \right)^d \exp\left(-\tau ||\beta||_1 \right) .$$

Show that the Lasso solution is the MAP (Maximum A Posteriori aka the mode of the posterior distribution) estimate of β, with $\lambda = 2\tau\sigma^2$.

Problem 5.
Suppose we estimate the regression coefficients in a linear regression model using an ℓ_1 penalty, that is find the minimiser of (1) with $p = 1$, for a particular value of λ. For questions *(a)* through *(e)*, indicate which of *(i)* through *(v)* is correct, and justify your answer.

(a) As we increase λ from 0, the training $RSS_1(\lambda)$ will:

(i) Increase initially, and then eventually start decreasing in an inverted U shape.

(ii) Decrease initially, and then eventually start increasing in a U shape.

(iii) Steadily increase.

(iv) Steadily decrease.

(b) Repeat *(a)* for test $RSS_1(\lambda)$.

(c) Repeat *(a)* for variance.

(d) Repeat *(a)* for squared bias.

Problem 6.
Recall that the lasso estimator $\hat{\beta}_\lambda$ satisfies (see page 25 of the lecture notes)

$$\mathbf{X}^t\mathbf{X}\hat{\beta}_\lambda = \mathbf{X}^t\mathbf{y} - \frac{\lambda}{2}\hat{z}\,,$$

for $\hat{z} \in \mathbb{R}^d$ such that $\hat{z} = \text{sign}(\hat{\beta}_\lambda)_j$ if $(\hat{\beta}_\lambda)_j \neq 0$ and $\hat{z} \in [-1,\, 1]$ if $(\hat{\beta}_\lambda)_j = 0$. Suppose in this problem that the columns of $\mathbf{X}$ are orthogonal.

(i) Argue that for $(\hat{\beta}_\lambda)_j \neq 0$, $\mathbf{X}^t\mathbf{y}$ and $(\hat{\beta}_\lambda)_j$ are of the same sign.

(ii) Deduce from *(i)* that in the orthogonal setting the lasso estimator is given by

$$(\hat{\beta}_\lambda)_j = \mathbf{X}_j^t\mathbf{y}\left(1 - \frac{\lambda}{2|\mathbf{X}_j^t\mathbf{y}|}\right)_+ .$$

Compare this expression with the one derived page 11 of the lecture notes.

Problem 7.
The lasso solution is unique when the rank$(\mathbf{X}) = d$. However, when rank$(\mathbf{X}) < d$, the criterion is not strictly convex, and there can be multiple minimisers of the lasso criterion.

(i) Show that if the lasso solution is not unique, then there exists uncountably many solutions.

(ii) Show that every lasso solution $\hat{\beta}_\lambda$ gives the same fitted value $\mathbf{X}\hat{\beta}_\lambda$.

(iii) Show that every lasso solution has the same ℓ_1 norm $||\hat{\beta}_\lambda||_1$.

Problem 8.
The elastic-net optimisation problem can be written as

$$\min_\beta \left\{||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda(\alpha||\beta||_2^2 + (1-\alpha)||\beta||_1)\right\}\,,$$

for some $0 \leq \alpha \leq 1$.

(i) Show how it is possible to turn this into a lasso problem, using an augmented version of $\mathbf{X}$ and $\mathbf{y}$. Specify the value of the tuning parameter in terms of λ and α.

(ii) Explain (using words) why an elastic net penalty enables feature selection, as well as coefficient shrinkage.

(iii) Write down a coordinate descent algorithm for the elastic-net optimisation problem.