SKETCH of the SOLUTIONS
STK 4030 - 2017

Exercise 1

a These plots show the strength of the shrinkage for the following three techniques as a function of the size of the correspondent parameter:

- left plot: variable selection (the estimate of the parameter is equal to 0 if its effect is non statistically different from 0, non-shrunk if it is statistically different from 0):
- centre plot: ridge regression (larger estimates are shrunk more);
- right plot: lasso (constant shrinkage, smaller estimates forced to be equal to 0).

b The two plots above explain lasso (left) and ridge (right) as a constrained regression methods. In particular, in the two-dimension problem showed, the solution for lasso must be within a square ($|\beta_1|+|\beta_2| \leq t$), that for ridge regression within the sphere ($\beta_1^2 + \beta_2^2 \leq t^2$). The former force some estimates of the parameters to be equal to 0, the latter does not.

c

$$\begin{aligned} E[\hat{\beta}_{\text{ridge}}] &= E[(X^T X + \lambda I_p)^{-1} X^T y] \\ &= E[(I_p + \lambda (X^T X)^{-1})^{-1} (X^T X)^{-1} X^T y] \\ &= W_\lambda E[\hat{\beta}_{\text{OLS}}] \end{aligned}$$

Since $\hat{\beta}_{\text{OLS}}$ is unbiased, $\hat{\beta}_{\text{ridge}}$ has larger bias but in the case $W_\lambda = I_p$, i.e., if and only if $\lambda = 0$.

d The variance of $\hat{\beta}_{\text{ridge}}$ is

$$\begin{aligned} \text{Var}[\hat{\beta}_{\text{ridge}}] &= \text{Var}[W_\lambda \hat{\beta}_{\text{OLS}}] \\ &= W_\lambda \text{Var}[\hat{\beta}_{\text{OLS}}] W_\lambda^T \\ &= \sigma^2 W_\lambda (X^T X)^{-1} W_\lambda^T \end{aligned}$$

and

$$\begin{aligned} \text{Var}[\hat{\beta}_{\text{OLS}}] - \text{Var}[\hat{\beta}_{\text{ridge}}] =& \sigma^2 \left[(X^T X)^{-1} - W_\lambda (X^T X)^{-1} W_\lambda \right] \\ =& \sigma^2 W_\lambda \left[(I_p + \lambda (X^T X)^{-1}) (X^T X)^{-1} (I_p + \right. \\ & \left. + \lambda (X^T X)^{-1})^T - (X^T X)^{-1} \right] W_\lambda^T \\ =& \sigma^2 W_\lambda \left[2\lambda (X^T X)^{-2} + \lambda^2 (X^T X)^{-3} \right] W_\lambda^T > 0 \end{aligned}$$

Exercise 2

a The procedure is incorrect because it does not use completely independent samples to build the model (including the choice of the tuning parameters) and to estimate the prediction error. In the first step, in particular, the predictors are selected using all the samples.

b Dimension reduction (the first step), model construction and choice of the tuning parameter should be performed only on the training part (K-1 folds) of the cross-validation split. Alternatively, a split in training and test sets should be made at the beginning, and the aforementioned steps only performed on it. In this case, the prediction error should be computed only on the test set.

Exercise 3

a The original idea of AdaBoost is to combine the results of a weak classifier applied to modifications of the data in order to obtain a good classification. In particular, the modification is performed by iteratively applying a weighting scheme in which more weight is assigned to the observations that are miss-classified in the previous iteration.

b We can rewrite the equation provided by the exercise,

$$(\alpha_m, G_m) = \text{argmin}_{\alpha,G} \sum_{i=1}^{N} \exp\{-y_i[f_{m-1}(x_i) + \frac{\alpha}{2}G(x_i)]\},$$

as

$$\begin{aligned}(\alpha_m, G_m) =&\text{argmin}_{\alpha,G} \sum_{i=1}^{N} \exp\{-y_i[\sum_{k=1}^{m-1} \frac{\alpha_k}{2}\hat{G}^{[k]}(x_i) + \frac{\alpha}{2}G(x_i)]\}\\ =&\text{argmin}_{\alpha,G} \sum_{i=1}^{N} w_i^{(m)} \exp\{y_i\frac{\alpha}{2}G(x_i)]\},\end{aligned}$$

where

$$w_i^{(m)} = \exp\{-y_i\hat{f}_{m-1}(x_i)\} = \exp\{-y_i \sum_{k=1}^{m-1} \frac{\alpha_k}{2}\hat{G}^{[k]}(x_i)\}.$$

Then, focusing on G,

$$\begin{aligned}
G_m =&\text{argmin}_G \sum_{i=1}^N w_i^{(m)} \exp\{y_i \frac{\alpha}{2} G(x_i)]\} \\
=&\text{argmin}_G \left(\sum_{G(x_i)=y_i} w_i^{(m)} \exp\{-\frac{\alpha}{2}\} + \sum_{G(x_i)\neq y_i} w_i^{(m)} \exp\{\frac{\alpha}{2}\} \right) \\
=&\text{argmin}_G \left(\exp\{-\frac{\alpha}{2}\} \sum_{i=1}^N w_i^{(m)} + \right. \\
&\left. + \left(\exp\{\frac{\alpha}{2}\} - \exp\{-\frac{\alpha}{2}\}\right) \sum_{G(x_i)\neq y_i} w_i^{(m)} \right) \\
=&\text{argmin}_G \left(\exp\{-\frac{\alpha}{2}\} \sum_{i=1}^N w_i^{(m)} + \right. \\
&\left. + \left(\exp\{\frac{\alpha}{2}\} - \exp\{-\frac{\alpha}{2}\}\right) \sum_{i=1}^N w_i^{(m)} I(G(x_i) \neq y_i) \right)
\end{aligned}$$

so $G_m = \text{argmin}_G \sum_{i=1}^N w_i^{(m)} I(y_i \neq G(x_i))$. For the explanation of the steps, see lecture 11 notes (page 4).

c Focusing on α, instead,

$$\alpha_m = \text{argmin}_\alpha \sum_{i=1}^N w_i^{(m)} \exp\{y_i \frac{\alpha}{2} G(x_i)]\}$$

Deriving with respect to α

$$\begin{aligned}
&- \sum_{G(x_i)=y_i} w_i^{(m)} \exp\{-\frac{\alpha}{2}\} + \sum_{G(x_i)\neq y_i} w_i^{(m)} \exp\{\frac{\alpha}{2}\} = 0 \\
&- \sum_{G(x_i)=y_i} w_i^{(m)} + \sum_{G(x_i)\neq y_i} w_i^{(m)} \exp\{\alpha\} = 0 \\
&\exp\{\alpha\} \frac{\sum_{G(x_i)\neq y_i} w_i^{(m)}}{\sum_{i=1}^N w_i^{(m)}} = \frac{\sum_{G(x_i)= y_i} w_i^{(m)}}{\sum_{i=1}^N w_i^{(m)}}
\end{aligned}$$

and

$$\alpha_m = \log\left(\frac{1 - \text{err}_m}{\text{err}_m}\right)$$

where

$$\text{err}_m = \frac{\sum_{i=1}^N w_i^{(m)} I(y_i \neq G_m(x_i))}{\sum_{i=1}^N w_i^{(m)}}$$

d Since

$$\hat{f}_m = \hat{f}_{m-1} + \alpha_m \hat{G}_m$$

then

$$w_i^{(m+1)} = w_i^{(m)} \exp\{-\frac{\alpha_m}{2} y_i G_m(x_i))\}$$

Using the fact that

$$-y_i G_m(x_i)) = 2I(G(x_i) \neq y_i) - 1,$$

the result follow.