Ridge Regression for an Affine Function

It is easy to adapt the above method to learn an affine function $f(x) = x^\top w + b$ instead of a linear function $f(x) = x^\top w$, where $b \in \mathbb{R}$. We have the following optimization program

\textbf{Program (RR3)}:

\begin{align*}
\text{minimize} & \quad \xi^\top \xi + Kw^\top w \\
\text{subject to} & \quad y - Xw - b1 = \xi,
\end{align*}

with $y, \xi, 1 \in \mathbb{R}^m$ and $w \in \mathbb{R}^n$. Note that in Program (RR3) minimization is performed over $\xi$, $w$ and $b$, but $b$ is \textit{not} penalized in the objective function.
Ridge Regression: Program (RR3) Solution

The objective function is \textit{convex}.

The Lagrangian associated with this program is

\[
L(\xi, w, b, \lambda) = \xi^\top \xi + Kw^\top w - w^\top X^\top \lambda - \xi^\top \lambda - b1^\top \lambda + \lambda^\top y.
\]

Since $L$ is \textit{convex as a function of $\xi, b, w$}, it has a minimum iff $\nabla L_{\xi,b,w} = 0$. 
Ridge Regression: Dual Function of (RR3)

We get

\[ \lambda = 2\xi \]
\[ 1^T \lambda = 0 \]
\[ w = \frac{1}{2K}X^T \lambda = X^T \frac{\xi}{K}. \]

As before, if we set \( \xi = K\alpha \), we obtain \( \lambda = 2K\alpha \), \( w = X^T\alpha \), and

\[ G(\alpha) = -K\alpha^T(XX^T + KI_m)\alpha + 2K\alpha^Ty. \]
Ridge Regression: Dual Program of (RR3)

Since $K > 0$ and $\lambda = 2K\alpha$, the dual to ridge regression is the following program

**Program (DRR3):**

minimize $\alpha^\top(XX^\top + KI_m)\alpha - 2\alpha^\top y$

subject to

$1^\top \alpha = 0$,

where the minimization is over $\alpha$. 
Ridge Regression: Solution to (DRR3)

Observe that up to the factor $1/2$, this problem satisfies the conditions of a previous proposition from the first lesson of the quadratic optimization lesson with

\[ A = (XX^T + KL_m)^{-1} \]
\[ b = y \]
\[ B = 1_m \]
\[ f = 0, \]

and $x$ renamed as $\alpha$. 
Ridge Regression: Solution to \((DRR3)\)

Therefore, it has a unique solution \((\alpha, \mu)\) (beware that \(\lambda = 2K\alpha\) is \textbf{not} the \(\lambda\) used before, which we rename as \(\mu\)), which is the unique solution of the KKT-equations

\[
\begin{pmatrix}
XX^T + Kl_m & 1_m \\
1_m^T & 0
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\mu
\end{pmatrix}
=
\begin{pmatrix}
y \\
0
\end{pmatrix}.
\]
Ridge Regression: Solution to (DRR3)

Since the solution is

$$\mu = (B^\top AB)^{-1}(B^\top Ab - f), \quad \alpha = A(b - B\mu),$$

we get

$$\mu = (1^\top(XX^\top + Kl_m)^{-1}1)^{-1}1^\top(XX^\top + Kl_m)^{-1}y$$
$$\alpha = (XX^\top + Kl_m)^{-1}(y - \mu 1).$$
Ridge Regression: Solution to (DRR3)

Interestingly $b = \mu$, which is not obvious a priori.

**Proposition.** We have $b = \mu$. 
In summary the KKT-equations determine both $\alpha$ and $\mu$, and so $w = X^T \alpha$ and $b$ as well.
Ridge Regression: Averaging Formula for $b$

There is also a useful expression of $b$ as an average. We have

$$b = \bar{y} - \sum_{j=1}^{n} \bar{X}^j w_j = \bar{y} - (\bar{X}^1 \cdots \bar{X}^n) w,$$

where $\bar{y}$ is the mean of $y$ and $\bar{X}^j$ is the mean of the $j$th column of $X$. 
Ridge Regression: Affine Case Reduction

It can be shown that solving the Dual (DRR3) for $\alpha$ and obtaining $w = X^T \alpha$ is equivalent to solving our previous ridge regression Problem (RR2) applied to the centered data $\hat{y} = y - \bar{y}1_m$ and $\hat{X} = X - \bar{X}$, where $\bar{X}$ is the $m \times n$ matrix whose $j$th column is $\bar{X}^j 1_m$, the vector whose coordinates are all equal to the mean $\bar{X}^j$ of the $j$th column $X^j$ of $X$. 
Ridge Regression: Program (RR6)

Program (RR6) is equivalent to ridge regression without an intercept term applied to the centered data $\hat{y} = y - \bar{y}1$ and $\hat{X} = X - \bar{X}$,

**Program (RR6):**

$$\begin{align*}
\text{minimize} & \quad \xi^T \xi + Kw^T w \\
\text{subject to} & \quad \hat{y} - \hat{X}w = \xi,
\end{align*}$$

minimizing over $\xi$ and $w$. 
Ridge Regression: Program (RR6) Solution

If $\hat{w}$ is the optimal solution of this program given by

$$\hat{w} = \hat{X}^T (\hat{X}\hat{X}^T + K_l)_{-1} \hat{y},$$

then $b$ is given by

$$b = \bar{y} - (\overline{X_1} \cdots \overline{X_n}) \hat{w}. $$
Ridge Regression: Learning an Affine Function

In practice Program (RR6) involving the centered data appears to be the preferred one.
Ridge Regression: Illustrated Example

Example. Consider the data set \((X, y_1)\) with

\[
X = \begin{pmatrix}
-10 & 11 \\
-6 & 5 \\
-2 & 4 \\
0 & 0 \\
1 & 2 \\
2 & -5 \\
6 & -4 \\
10 & -6
\end{pmatrix}, \quad y_1 = \begin{pmatrix}
0 \\
-2.5 \\
0.5 \\
-2 \\
2.5 \\
-4.2 \\
1 \\
4
\end{pmatrix}
\]

as illustrated in Figure 1.
Ridge Regression: Illustrated Example

We find that $\bar{y} = -0.0875$ and $(\bar{X}^1, \bar{X}^2) = (0.125, 0.875)$. For the value $K = 5$, we obtain

$$w = \begin{pmatrix} 0.9207 \\ 0.8677 \end{pmatrix}, \quad b = -0.9618,$$

for $K = 0.1$, we obtain

$$w = \begin{pmatrix} 1.1651 \\ 1.1341 \end{pmatrix}, \quad b = -1.2255,$$

and for $K = 0.01$,

$$w = \begin{pmatrix} 1.1709 \\ 1.1405 \end{pmatrix}, \quad b = -1.2318.$$

See Figure 2.
Ridge Regression: Illustrated Example

Figure 1: The data set $(X, y_1)$. 
Figure 2: The graph of the plane $f(x, y) = 1.1709x + 1.1405y - 1.2318$ as an approximate fit to the data $(X, y_1)$. 
Ridge Regression: Illustrated Example

We conclude that the points \((X_i, y_i)\) (where \(X_i\) is the \(i\)th row of \(X\)) almost lie on the plane of equation

\[ x + y - z - 1 = 0, \]

and that \(f\) is almost the function given by \(f(x, y) = 1.1x + 1.1y - 1.2\). See Figures 3 and 4.
Ridge Regression: Illustrated Example

Figure 3: The graph of the plane \( f(x, y) = 1.1x + 1.1y - 1.2 \) as an approximate fit to the data \((X, y_1)\).
Figure 4: A comparison of how the graphs of the planes corresponding to $K = 1, 0.1, 0.01$ and the salmon plane of equation $f(x, y) = 1.1x + 1.1y - 1.2$ approximate the data $(X, y_1)$. 
Ridge Regression: Illustrated Example

If we change $y_1$ to

$$y_2 = \begin{pmatrix} 0 & -2 & 1 & -1 & 2 & -4 & 1 & 3 \end{pmatrix}^\top,$$

as evidenced by Figure 5, the exact solution is

$$w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad b = -1,$$

and for $K = 0.01$, we find that

$$w = \begin{pmatrix} 0.9999 \\ 0.9999 \end{pmatrix}, \quad b = -0.9999.$$
Ridge Regression: Illustrated Example

Figure 5: The data \((X, y_2)\) is contained within the graph of the plane \(f(x, y) = x + y - 1\).
Ridge Regression: Learning an Affine Function

We can see how the choice of $K$ affects the quality of the solution $(w, b)$ by computing the norm $\|\xi\|_2$ of the error vector $\xi = \hat{y} - \hat{X}w$. We notice that the smaller $K$ is, the smaller is this norm.

As a least squares problem, the solution is given in terms of the pseudo-inverse $[X 1]^+$ of $[X 1]$ by

$$\begin{pmatrix} w \\ b \end{pmatrix} = [X 1]^+ y.$$