The above suggests defining the variable α so that $\xi = K\alpha$, so we have $\lambda = 2K\alpha$ and $w = X^\top \alpha$. Then we obtain the dual function as a function of α by substituting the above values of ξ, λ and w back in the Lagrangian and we get

$$\begin{aligned}
G(\alpha) &= K^2\alpha^\top\alpha + K\alpha^\top XX^\top\alpha - 2K\alpha^\top XX^\top\alpha - 2K^2\alpha^\top\alpha + 2K\alpha^\top y \\
&= -K\alpha^\top(XX^\top + KI_m)\alpha + 2K\alpha^\top y.
\end{aligned}$$

This is a strictly concave function so by Theorem 40.13(4), its maximum is achieved iff $\nabla G_\alpha = 0$, that is,

$$2K(XX^\top + KI_m)\alpha = 2Ky,$$

which yields

$$\alpha = (XX^\top + KI_m)^{-1}y.$$

Putting everything together we obtain

$$\begin{aligned}
\alpha &= (XX^\top + KI_m)^{-1}y \\
w &= X^\top\alpha \\
\xi &= K\alpha,
\end{aligned}$$

which yields

$$w = X^\top(XX^\top + KI_m)^{-1}y. \tag{$*_{wd}$}$$

Earlier in $(*_{wp})$ we found that

$$w = (X^\top X + KI_n)^{-1}X^\top y,$$

and it is easy to check that

$$(X^\top X + KI_n)^{-1}X^\top = X^\top(XX^\top + KI_m)^{-1}.$$

If $n < m$ it is cheaper to use the formula on the left-hand side, but if $m < n$ it is cheaper to use the formula on the right-hand side.

55.2 Ridge Regression; Learning 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

Program (RR3):

$$\begin{aligned}
&\text{minimize} \quad \xi^\top\xi + Kw^\top w \\
&\text{subject to} \\
&\qquad\qquad y - Xw - b\mathbf{1} = \xi,
\end{aligned}$$