As in the standard case of ridge regression, if $F = \mathbb{R}^n$ (but the inner product $\langle -, - \rangle$ is arbitrary), we can adapt the above method to learn an affine function $f(w) = x^\top w + b$ instead of a linear function $f(w) = x^\top w$, where $b \in \mathbb{R}$. This time we assume that b is of the form

$$b = \overline{y} - \langle w, (\overline{X^1} \cdots \overline{X^n}) \rangle,$$

where X^j is the j column of the $m \times n$ matrix X whose ith row is the transpose of the column vector $\varphi(x_i)$, and where $(\overline{X^1} \cdots \overline{X^n})$ is viewed as a column vector. We have the minimization problem

Program (KRR6′):

$$\begin{aligned} &\text{minimize} \quad \xi^\top \xi + K\langle w, w \rangle \\ &\text{subject to} \\ &\qquad\qquad \widehat{y_i} - \langle w, \widehat{\varphi(x_i)} \rangle = \xi_i, \quad i = 1, \ldots, m, \end{aligned}$$

minimizing over ξ and w, where $\widehat{\varphi(x_i)}$ is the n-dimensional vector $\varphi(x_i) - (\overline{X^1} \cdots \overline{X^n})$.

The solution is given in terms of the matrix $\widehat{\mathbf{G}}$ defined by

$$\widehat{\mathbf{G}}_{ij} = \langle \widehat{\varphi(x_i)}, \widehat{\varphi(x_j)} \rangle,$$

as before. We get

$$\alpha = (\widehat{\mathbf{G}} + K I_m)^{-1} \widehat{y},$$

and according to a previous computation, b is given by

$$b = \overline{y} - \frac{1}{m} \mathbf{1} \widehat{\mathbf{G}} \alpha.$$

We explained in Section 53.4 how to compute the matrix $\widehat{\mathbf{G}}$ from the matrix $\mathbf{G}$.

Since the dimension of the feature space F may be very large, one might worry that computing the inner products $\langle \varphi(x_i), \varphi(x_j) \rangle$ might be very expensive. This is where kernel functions come to the rescue. A *kernel function* κ for an embedding $\varphi \colon \mathbb{R}^n \to F$ is a map $\kappa \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the property that

$$\kappa(u, v) = \langle \varphi(u), \varphi(v) \rangle \quad \text{for all } u, v \in \mathbb{R}^n.$$

If $\kappa(u, v)$ can be computed in a reasonably cheap way, and if $\varphi(u)$ can be computed cheaply, then the inner products $\langle \varphi(x_i), \varphi(x_j) \rangle$ (and $\langle \varphi(x_i), \varphi(x) \rangle$) can be computed cheaply; see Chapter 53. Fortunately there are good kernel functions. Two very good sources on kernel methods are Schölkopf and Smola [145] and Shawe–Taylor and Christianini [159].