the matrix $XX^\top$ consists of the inner products $x_i^\top x_j$, and similarly the function learned $f(x) = x^\top w$ can be expressed as

$$f(x) = \sum_{i=1}^{m} \alpha_i x_i^\top x,$$

namely that both w and $f(x)$ are given *in terms of the inner products* $x_i^\top x_j$ and $x_i^\top x$.

This fact is the key to a generalization to ridge regression in which the input space $\mathbb{R}^n$ is embedded in a larger (possibly infinite dimensional) Euclidean space F (with an inner product $\langle -, - \rangle$) usually called a *feature space*, using a function

$$\varphi\colon \mathbb{R}^n \to F.$$

The problem becomes (*kernel ridge regression*)

Program (KRR2):

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

minimizing over ξ and w. Note that $w \in F$. This problem is discussed in Shawe–Taylor and Christianini [159] (Section 7.3).

We will show below that the solution is exactly the same:

$$\begin{aligned} \alpha &= (\mathbf{G} + KI_m)^{-1} y \\ w &= \sum_{i=1}^{m} \alpha_i \varphi(x_i) \\ \xi &= K\alpha, \end{aligned}$$

where $\mathbf{G}$ is the Gram matrix given by $\mathbf{G}_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$. This matrix is also called the *kernel matrix* and is often denoted by $\mathbf{K}$ instead of $\mathbf{G}$.

In this framework we have to be a little careful in using gradients since the inner product $\langle -, - \rangle$ on F is involved and F could be infinite dimensional, but this causes no problem because we can use derivatives, and by Proposition 39.5 we have

$$d\langle -, - \rangle_{(u,v)}(x, y) = \langle x, v \rangle + \langle u, y \rangle.$$

This implies that the derivative of the map $u \mapsto \langle u, u \rangle$ is

$$d\langle -, - \rangle_u(x) = 2\langle x, u \rangle. \tag{d_1}$$