the best set (see Section 10.2 and Chapter 18). As an aside, note that in the case of Kernel PCA (see Section 1.7 and Chapter 14 below), one does not need to worry about the whitening step in (2.59) and (2.60): using the canonical dot product in $\mathbb{R}^m$ (rather than $\langle \cdot, \cdot \rangle$) will simply lead to diagonalizing K^2 instead of K, which yields the same eigenvectors with squared eigenvalues. This was pointed out by [350, 361]. The study [361] reports experiments where (2.56) was employed to speed up Kernel PCA by choosing $\{z_1, \ldots, z_n\}$ as a subset of $\{x_1, \ldots, x_m\}$.

2.2.7 A Kernel Map Defined from Pairwise Similarities

In practice, we are given a finite amount of data $x_1, \ldots, x_m$. The following simple observation shows that even if we do not want to (or are unable to) analyze a given kernel k analytically, we can still compute a map Φ such that k corresponds to a dot product in the linear span of the $\Phi(x_i)$:

Proposition 2.16 (Data-Dependent Kernel Map [467]) *Suppose the data $x_1, \ldots, x_m$ and the kernel k are such that the kernel Gram matrix $K_{ij} = k(x_i, x_j)$ is positive definite. Then it is possible to construct a map Φ into an m-dimensional feature space $\mathcal{H}$ such that*

$$k(x_i, x_j) = \left\langle \Phi(x_i), \Phi(x_j) \right\rangle. \tag{2.61}$$

Conversely, given an arbitrary map Φ into some feature space $\mathcal{H}$, the matrix $K_{ij} = \left\langle \Phi(x_i), \Phi(x_j) \right\rangle$ is positive definite.

Proof First assume that K is positive definite. In this case, it can be diagonalized as $K = SDS^\top$, with an orthogonal matrix S and a diagonal matrix D with nonnegative entries. Then

$$k(x_i, x_j) = (SDS^\top)_{ij} = \langle S_i, DS_j \rangle = \left\langle \sqrt{D} S_i, \sqrt{D} S_j \right\rangle, \tag{2.62}$$

where we have defined the S_i as the rows of S (note that the columns of S would be K's eigenvectors). Therefore, K is the Gram matrix of the vectors $\sqrt{D} \cdot S_i$.[9] Hence the following map Φ, defined on $x_1, \ldots, x_m$ will satisfy (2.61)

$$\Phi : x_i \mapsto \sqrt{D} \cdot S_i. \tag{2.63}$$

Thus far, Φ is only defined on a set of points, rather than on a vector space. Therefore, it makes no sense to ask whether it is linear. We can, however, ask whether it can be *extended* to a linear map, provided the x_i are elements of a vector space. The answer is that if the x_i are linearly dependent (which is often the case), then this will not be possible, since a linear map would then typically be over-

puted as $D_n^{-1/2} U_n^\top (k(z_1, x), \ldots, k(z_n, x))$, where $U_n D_n U_n^\top$ is the eigenvalue decomposition of K_n. Note that the columns of U_n are the eigenvectors of K_n. We discard all columns that correspond to zero eigenvalues, as well as the corresponding dimensions of D_n. To *approximate* the map, we may actually discard all eigenvalues smaller than some $\epsilon > 0$.

9. In fact, every positive definite matrix is the Gram matrix of some set of vectors [46].