Let us have two-dimensional vectors:

$$\mathbf{x} = (x_1, x_2), \quad \mathbf{z} = (z_1, z_2)$$

and you use a **polynomial kernel of degree 2**, which is:

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x} \cdot \mathbf{z} + 1)^2$$

Let's expand this:

First, compute the dot product:

$$\mathbf{x} \cdot \mathbf{z} = x_1 z_1 + x_2 z_2$$

Thus:

$$K(\mathbf{x}, \mathbf{z}) = (x_1 z_1 + x_2 z_2 + 1)^2$$

Expand the square:

$$= (x_1 z_1)^2 + 2(x_1 z_1)(x_2 z_2) + (x_2 z_2)^2 + 2(x_1 z_1) + 2(x_2 z_2) + 1$$

But we want to **find the explicit feature map** $\phi(\mathbf{x})$, so that:

$$K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x}) \cdot \phi(\mathbf{z})$$

The trick is to define $\phi(\mathbf{x})$ with all degree 2 monomials and degree 1 monomials (and constant term), namely:

$$\phi(\mathbf{x}) = \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2, \sqrt{2} x_1, \sqrt{2} x_2, 1 \right)$$

The square roots of 2 are there to match the expansion properly and make inner products work out exactly.

Summary:
The explicit **transfer function** (feature mapping) is:

$$\boxed{\phi(x_1, x_2) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2, \sqrt{2} x_1, \sqrt{2} x_2, 1)}$$

Representer Theorem and the Role of the Norm in Kernel Methods

In many machine learning problems, especially those involving kernels, we aim to find a function f that fits the training data well while remaining simple. This idea is formalized through **regularized risk minimization**, where we minimize an objective of the form:

$$\min_f \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2$$

Here:

- $L(y_i, f(x_i))$ measures how well f fits the data (loss function),
- $\|f\|_{\mathcal{H}}$ is the norm of f in a Hilbert space $\mathcal{H}$,
- $\lambda > 0$ is a regularization parameter controlling the trade-off between fitting and simplicity.

The **Representer Theorem** states that even though f could live in an infinite-dimensional space, the solution can always be written as a **finite linear combination** of the training examples mapped into the feature space:

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

where:

- $K(x_i, x)$ is the **kernel function**, which measures similarity between x_i and x,
- and α_i are real-valued coefficients to be determined.

The kernel function K implicitly defines a feature mapping ϕ such that:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

meaning $K(x_i, x_j)$ computes the inner product between x_i and x_j in the feature space, without needing to explicitly map them.

Thus, instead of searching over all possible functions, it suffices to find the n coefficients α_i, greatly simplifying the optimization.

Understanding the Norm $||f||_{\mathcal{H}}$

The term $||f||_{\mathcal{H}}$ measures the "size" or "complexity" of the function f in the Hilbert space $\mathcal{H}$.

- A **Hilbert space** is a space where notions like angles and lengths between functions make sense, much like vectors in Euclidean space.
- In kernel methods, the Hilbert space is the space induced by the kernel, where data may be mapped into very high or infinite dimensions.
- The **norm** $||f||_{\mathcal{H}}$ expresses how complex or "wiggly" f is in this space.

Minimizing $||f||_{\mathcal{H}}$ encourages f to be simple and smooth:

- In linear models with a linear kernel, $||f||_{\mathcal{H}}$ reduces to $||w||$, the standard Euclidean norm of the weight vector w.
- In Gaussian (RBF) kernels, minimizing $||f||_{\mathcal{H}}$ encourages functions that are smooth over the input space.

Thus, the regularization term $\lambda||f||^2_{\mathcal{H}}$ plays a critical role in controlling overfitting and ensuring that the learned function generalizes well to unseen data.