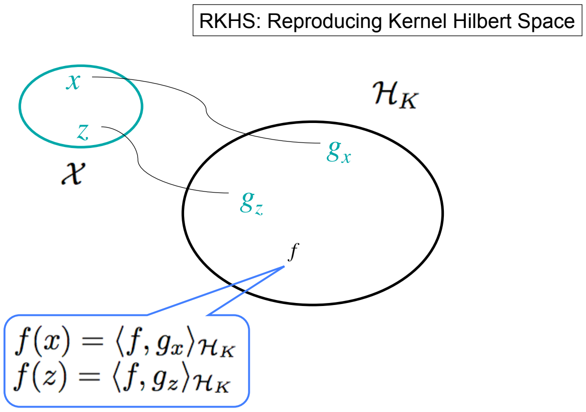
## Reproducing Kernel Hilbert Space (RKHS)

A reproducing Kernel Hilbert space $\mathcal{H}_K$ is a collection of functions $f(x)$ defined on domain $\mathcal{X}$ and equipped with an inner product $\langle f, g\rangle_{\mathcal{H}_K}$. (The meaning of $K$ in the subscript will be explained soon.) What's special about RKHS is that for each $x$ in the domain $\mathcal{X}$, there exists a function in $\mathcal{H}_K$, denoted by $g_x$, such that for any function in this RKHS, its value at $x$ can be obtained by doing an inner product with $g_x$, known as the *reproducing property*.


RKHS: Reproducing Kernel Hilbert Space
$x$
$z$
$\mathcal{X}$
$\mathcal{H}_K$
$g_x$
$g_z$
$f$
$f(x) = \langle f, g_x\rangle_{\mathcal{H}_K}$
$f(z) = \langle f, g_z\rangle_{\mathcal{H}_K}$

Define the following bivariate function $K(\cdot,\cdot)$ on $\mathcal{X}\times\mathcal{X}$:

$$K(x,z) = \langle g_x, g_z\rangle_{\mathcal{H}_K}.$$

It is easy to show that $K(x,z)$ is symmetric and psd (positive semidefinite) using the property of inner products.

$$\begin{aligned} K(x,z) &= \langle g_x, g_z\rangle_{\mathcal{H}_K} = \langle g_z, g_x\rangle_{\mathcal{H}_K} \\ &= K(z,x); \quad \text{inner product is symmetric} \end{aligned}$$

$$\begin{aligned} \sum_{i,j}\alpha_i\alpha_j K(x_i,x_j) &= \sum_{i,j}\alpha_i\alpha_j\langle g_{x_i}, g_{x_j}\rangle_{\mathcal{H}_K} = \Big\langle \sum_i \alpha_i g_{x_i}, \sum_j \alpha_j g_{x_j}\Big\rangle_{\mathcal{H}_K} \\ &= \Big\|\sum_i \alpha_i g_{x_i}\Big\|^2 \geq 0. \end{aligned}$$

## A Simple Example of RKHS

Does such a function space exist? Let's look at a simple example of RKHS, which contains all linear functions (without intercept) on $\mathcal{X} = \mathbb{R}^2$:

$$\mathcal{H}_K = \left\{ f_\beta(x) = \beta^t x = \beta_1 x_1 + \beta_2 x_2, \beta \in \mathbf{R}^2 \right\}$$

with the inner product between two linear functions, $f_\beta$ and $f_\theta$, being the ordinary dot product between their slope coefficient vectors, namely,

$$\langle f_\beta, f_\theta \rangle_{\mathcal{H}_K} = \beta \cdot \theta = \beta^t \theta.$$

- The reproducing property: for any $z \in \mathbf{R}^2$,

  $$f_\beta(z) = \beta^t z = \langle f_\beta, f_z \rangle,$$

  where $f_z$ is a linear function with slope vector $(z_1, z_2)^t$. For example, for any $f_\beta \in \mathcal{H}_K$, its value at $z = (2, 3)^t$ is equal to

  $$f_\beta(2, 3) = \beta_1(2) + \beta_2(3) = \langle f_\beta, g \rangle,$$

  where $g(x) = 2x_1 + 3x_2$.

- The kernel function

  $$K(x, z) = \langle f_x, f_z \rangle = x^t z$$

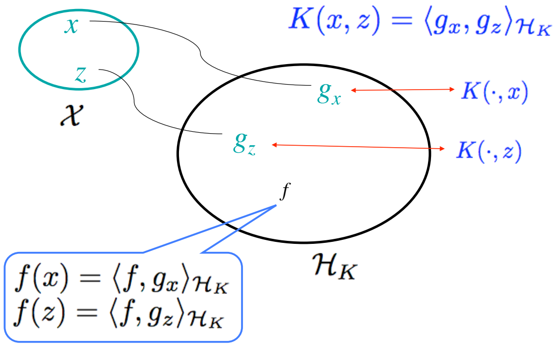
  is the dot product in $\mathbf{R}^2$. Note that if we fix one argument of $K$, then $K(\cdot, z)$ is an element in $\mathcal{H}_K$ and it's basically the evaluation function $f_z$: $\langle f_\beta, K(\cdot, z) \rangle = f_\beta(z)$.

## The Kernel Function

It turns out that a symmetric and psd bivariate kernel function $K(\cdot, \cdot)$ uniquely determines an RKHS. This is why when we talk about an RKHS, we do not need to explain what that space looks like; we just need to give the expression of the $K$ function.

Given an RKSH, we can define a kernel function $K(\cdot,\cdot)$ that is symmetric and psd. Next we show that given a symmetric and psd bivariate function $K(\cdot,\cdot)$, we can construct an RKHS. The construction involves the following steps:

- Note that if we fix one argument of $K$, then $K(\cdot, z)$ is a function defined on $\mathcal{X}$. Construct a function space $\mathcal{H}$ that contains all such functions $K(\cdot, z)$, their linear combinations, and their limits:
$$\mathcal{H} = \overline{\text{span}\{K(\cdot, z), z \in \mathcal{X}\}}.$$
- Next define an inner product between two elements in $\mathcal{H}$ to make $\mathcal{H}$ a Hilbert space. First, define
$$\langle K(\cdot, z), K(\cdot, x)\rangle_{\mathcal{H}} \triangleq K(x, z).$$
Then for $f(x) = \sum_k a_k K(\cdot, z_k)$ and $g(x) = \sum_j b_j K(\cdot, x_j)$, define
$$\langle f, g\rangle_{\mathcal{H}} = \sum_k \sum_j a_k b_j K(x_j, z_k). \tag{1}$$
Since $K$ is symmetric and psd, the inner product $\langle \cdot,\cdot\rangle_{\mathcal{H}}$ is well-defined. Further, it is easy to check that $K(\cdot, z)$ plays the same role as the $g_z$ function, i.e., for any $f \in \mathcal{H}$, its value at $z$ can be reproduced by computing its inner product with $K(\cdot, z)$:
$$f(z) = \langle f, K(\cdot, z)\rangle_{\mathcal{H}}.$$


$K(x,z) = \langle g_x, g_z\rangle_{\mathcal{H}_K}$
$x$
$z$
$\mathcal{X}$
$g_x$
$K(\cdot,x)$
$g_z$
$K(\cdot,z)$
$f$
$\mathcal{H}_K$
$f(x) = \langle f, g_x\rangle_{\mathcal{H}_K}$
$f(z) = \langle f, g_z\rangle_{\mathcal{H}_K}$

- Note that we haven't checked whether $\{K(\cdot, z), z \in \mathcal{X}\}$ form a basis for $\mathcal{H}$; actually they do NOT. So the linear representation for $f$ and the one for $g$ are not unique. For example, we could write the same $f$ function as $f(x) = \sum_l c_l K(\cdot, u_l)$. So we need to show that the inner product defined at (1) stays the same if we plug in a different representation of $f$ or $g$. Further, we need to show that the inner product (1) is well-defined when $f$ or $g$ are limits of linear combinations of $K(\cdot, z)$'s.

  I skip the proof here. For details, you can google lecture notes or review papers on RKHS, or check this book "Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond" by Schlkopf and Smola.

## Representer Theorem

Consider the following function estimation problem: given a set of training data $(x_i, y_i)_{i=1}^n$, find a function from an RKHS $\mathcal{H}_K$ that minimizes the following "loss + penalty" objective function:

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

The first term denotes the empirical loss on the $n$ sample points ($y_i$ could be continuous as in regression and categorical as in classification), the second term is generally interpreted as a roughness penalty, where the roughness is measured by the squared norm $\|f\|^2_{\mathcal{H}_K}$, and $\lambda$ is a tuning parameter.

The RKHS could be an infinite dimensional function space. A beautiful result, known as the *Representer Theorem* (Kimeldorf and Wahba, 1971), shows that the minimizer of (2) is always finite dimensional (with maximal dim $= n$) and takes the following form

$$\begin{aligned} & \text{argmin}_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^n L(y_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_K} \\ = \;& w_1 K(x, x_1) + \cdots + w_n K(x, x_n). \end{aligned}$$

**Proof :** Let $\mathcal{H}_1 = \text{span}\{K(\cdot, x_1), \ldots, K(\cdot, x_n)\}$ and $\mathcal{H}_2 = \mathcal{H}_1^{\perp}$. Then for any function $f \in \mathcal{H}_K$, we can write

$$f = f_1 + f_2, \quad \text{where } f_1 \in \mathcal{H}_1 \text{ and } f_2 \in \mathcal{H}_2.$$

Then we have the following

1. $\|f\|^2 \geq \|f_1\|^2$;
2. $f(x_i) = f_1(x_i)$ for $i = 1, \ldots, n$, because

$$\big\langle f, K(\cdot, x_i) \big\rangle_{\mathcal{H}_K} = \big\langle f_1 + f_2, K(\cdot, x_i) \big\rangle_{\mathcal{H}_K} = \big\langle f_1, K(\cdot, x_i) \big\rangle_{\mathcal{H}_K}.$$

That is $\Omega(f) \geq \Omega(f_1)$. So to minimize $\Omega(f)$, it suffices to focus on subspace $\mathcal{H}_1$. (Does the proof sound familiar? Yes, it follows the same argument as the one in the proof for smoothing splines.)