In discussing identification of systems, we spoke of fitting linear models (e.g. for indirect adaptive control). More generally one might speak of learning a (nonlinear) map $f: X \rightarrow Y$ from a set $X$ of stimuli/input to response/output set $Y$, using an empirical data sequence \( \{ (x_i, y_i) : i = 1, 2, \ldots, m \} \). The index set \{ $i = 1, 2, \ldots, m$ \} need not represent time instants. We refer to this setting as supervised learning or "learning with a teacher". One widely used approach to this task is outlined here and is known as kernel-based learning.

Some desiderata:

(i) Learned model $\hat{f}$ should account for the data by approximating it well in some sense.

(ii) Learned model $\hat{f}$ should prove effective in making generalizations/predictions, i.e. error $\| y - \hat{f}(x) \|$ between prediction $\hat{f}(x)$ and observed response $y$ to a future stimulus $x$ should be small.

(iii) For the most part, $X \subset \mathbb{R}^n$ is a closed set and $Y = \mathbb{R}^1$. 
Various formulations of reasoning show that these are competing requirements. Models of high complexity that fit the data well pay a price in generalization performance. This trade-off is encoded as an optimization problem on a suitable space of candidate models, known as a hypothesis space. The very definition of this space is based on the concept of a kernel and associated Hilbert space known as reproducing kernel Hilbert space (RKHS). Here we go with mathematical details.

A Hilbert space \( V \) over reals \( \mathbb{R} \) is a vector space with a positive definite inner product \( \langle \cdot, \cdot \rangle \), and associated norm satisfying completeness property. Thus

\[
\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}
\]

satisfies

(i) \( \langle v, w \rangle = \langle w, v \rangle \)

(ii) \( \langle \alpha v, w \rangle = \alpha \langle v, w \rangle \quad \alpha \in \mathbb{R} \)

(iii) \( \langle v_1 + v_2, w \rangle = \langle v_1, w \rangle + \langle v_2, w \rangle \)

(iv) \( \langle v, v \rangle > 0 \) and \( \langle v, v \rangle = 0 \Rightarrow v = 0 \)

Norm \( \|v\| = (\langle v, v \rangle)^{1/2} \) defines a metric

\[
d(v, w) = \|v - w\|
\]
We say a sequence \( \{ u_n : n = 1, 2, 3, \ldots \} \subseteq V \) converges to \( v \in V \) if
\[
\lim_{n \to \infty} \| u_n - v \| = 0.
\]

We say a sequence \( \{ u_n : n = 1, 2, 3, \ldots \} \subseteq V \) is a Cauchy sequence if
\[
\lim_{n, m \to \infty} \| u_n - u_m \| = 0.
\]

It is easy to see that every convergent sequence is a Cauchy sequence. In general the converse is not true. We say that \((V, \| \cdot \|)\) is a complete normed linear space if every Cauchy sequence is also a convergent sequence.

**Example.** Suppose \( X = \{1, 2, \ldots, L\} \subseteq \mathbb{R}^1 \), a discrete set of stimuli inputs. Each \( f : X \to \mathbb{R}^L \) defines a row vector \((f(1), f(2), \ldots, f(L)) \in \mathbb{R}^L\). A kernel \( K \) is simply a function
\[
K : X \times X \to \mathbb{R}
\]
\[
(i, j) \mapsto K(i, j)
\]
satisfying

(i) \( K(i, j) = K(j, i) \) (symmetry)

(ii) For any real \( c_i, \) \( i = 1, 2, \ldots, L \)
\[
\sum_{i=1}^{L} \sum_{j=1}^{L} c_i c_j K(i, j) \geq 0
\]
Condition (ii) above is often referred to as **positive definiteness** of the kernel function, but in linear algebra this would correspond to **positive semidefiniteness** of the matrix

\[ K = \begin{bmatrix} K(i, j) \end{bmatrix} \]

with \( L \) rows and \( L \) columns.

Define the set of functions

\[ k_i : X \rightarrow \mathbb{R} \]

\[ j \mapsto k_i(j) = K(i, j) \]

Again each such \( k_i \) defines a row vector

\( (k_i(1), k_i(2), \ldots, k_i(L)) \in \mathbb{R}^L \). Thus it makes sense to look for a model \( f : X \rightarrow \mathbb{R} \) in the hypothesis space

\[ H_K = \left\{ f : X \rightarrow \mathbb{R} \mid f = \sum_{i=1}^{L} a_i k_i, a_i \in \mathbb{R} \right\} \]

Clearly this is at most \( L \) dimensional.

In fact, for any \( f \in H_K \), associated row vector

\[
(f(1) \ f(2) \ \ldots \ f(L)) = (a_1 \ a_2 \ \ldots \ a_L) \begin{pmatrix} k_1(1) & k_1(2) & \ldots & k_1(L) \\ k_2(1) & k_2(2) & \ldots & k_2(L) \\ \vdots & \vdots & \ddots & \vdots \\ k_L(1) & k_L(2) & \ldots & k_L(L) \end{pmatrix}
\]
or in short hand, the row vector

\[ f = a \mathbf{1}_k \]

where \( a = (a_1, a_2, \ldots, a_L) \) and
\n\[ \text{dim}(\mathbf{H}_k) = \text{dim}(\text{range}(\mathbf{K})) \]

On \( \mathbf{H}_k \) define inner product candidate

\[ \langle f, g \rangle_k = \langle \sum_{i=1}^{L} a_i \mathbf{k}_i, \sum_{j=1}^{L} b_j \mathbf{k}_j \rangle \]

\[ = \sum_{i=1}^{L} \sum_{j=1}^{L} a_i b_j \langle \mathbf{k}_i, \mathbf{k}_j \rangle_k \]

where \( \langle \mathbf{k}_i, \mathbf{k}_j \rangle_k = K(i, j) \).

By positive definiteness of the kernel \( K \),
(hence positive semidefiniteness of the matrix \( \mathbf{K} \)),

\[ \langle f, f \rangle_k = \sum_{i=1}^{L} \sum_{j=1}^{L} a_i a_j K(i, j) \geq 0 \]

The r.h.s above can be written as \( a \mathbf{K} a^T \)
where \( a \) is a row vector and \( a^T \) is associated column vector. Since \( \mathbf{K} \) is positive semidefinite
\[ a = (a_1, a_2, \ldots, a_L) \]

**Proof**

Let \( f = (f_1, f_2, \ldots, f_L) \) for some row vector, and let \( f(c_i) = \langle f, \mathbf{k} \rangle \mathbf{k} \).

In \( K \), \( \langle f, \mathbf{k} \rangle \) is a reproducing kernel Hilbert space

It follows that \( (K, \langle \cdot, \cdot \rangle) \) is a genuine inner product.

Then \( \langle f, f \rangle = 0 \) if and only if \( a_k = 0 \).

It can be factorized as \( \mathbf{K} = \mathbf{N} \mathbf{T} \).
\[ f(i) = \left( \sum_{j=1}^{L} a_j \cdot K_j \right)(i) \]
\[ = \sum_{j=1}^{L} a_j \cdot K_j(i) \]
\[ = \sum_{j=1}^{L} a_j \cdot K(i, i) \]

\[ \langle f, K_i \rangle_K = \langle \sum_{j=1}^{L} a_j \cdot K_j, K_i \rangle_K \]
\[ = \sum_{j=1}^{L} a_j \cdot \langle K_j, K_i \rangle_K \]
\[ = \sum_{j=1}^{L} a_j \cdot K(i, i) \]

Hence \[ f(i) = \langle f, K_i \rangle_K \]
for \( i = 1, 2, \ldots, L \).

We can proceed from the example above of a discrete, finite set of stimuli/inputs to the more general setting of a closed possibly a continuum and kernel
\[ K : X \times X \to \mathbb{R} \]
satisfying

\[(\text{symmetry}) \quad K(x, x') = K(x', x), \quad x, x' \in X\]

and, for any choice of \(c_1, c_2, \ldots, c_m \in \mathbb{R}\),
and, for any choice of \(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m \in X\)
and, for any positive integer \(m\),

\[\sum_{i=1}^{m} \sum_{j=1}^{m} c_i c_j K(\tilde{x}_i, \tilde{x}_j) \geq 0.\]

For now we are SILENT on continuity of \(K\) etc.

Associated to such a positive definite kernel \(K_\cdot\) we define kernel functions, \(K_{\tilde{x}_\cdot}\) for any \(\tilde{x} \in X\) by letting

\[K_{\tilde{x}} : X \to \mathbb{R}\]

\[x \mapsto K_\tilde{x}(x) = K(\tilde{x}, x)\]

Notice that when \(X\) is a continuum, the family of kernel functions is uncountably infinite.

Define

\[H_k^{pre} = \left\{ f : X \to \mathbb{R} \mid f = \sum_{i=1}^{m} a_i K_{\tilde{x}_i}, \quad m \text{ any positive integer} \right\}\]

\[\forall a_i \in \mathbb{R}, \forall \tilde{x}_i \in X\]
$H_k \text{ is a (in general infinite dimensional) vector space.}$

First define $\langle K_{x_i}, K_{x_j} \rangle_k = K(x_i, x_j)$ a positive definite inner product on all of $H^\text{pre}_k$.

and extend this by linearity to give a definite inner product on all of $H^\text{pre}_k$.

Next, define $\|f\|_k = \sqrt{\langle f, f \rangle_k}$ as norm on $H^\text{pre}_k$. Then, our hypothesis space

$H_k = \text{completion of } H^\text{pre}_k$

( hypotheses about $K$ such as continuity matter here )

in the above norm, and one can verify that the inner product on $H^\text{pre}_k$ extends to one on $H_k$, satisfying the reproducing property,

$F(x) = \langle f, K_x \rangle_k$

We are now ready to discuss the modeling problem:

Given a set of input-output data

$\{(x_i, y_i) : i = 1, 2, \ldots, m\} \subset X \times Y$, find $f \in H_k$ such that

$C(f) = \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i, f(x_i)) + \lambda \|f\|_k^2$.
Here $\Phi$ is a loss function measuring fit error, e.g. $\Phi(a, b) = (a - b)^2$.

$\|f\|^2_2$ is a measure of complexity of $f$.

The constant $\lambda > 0$ is chosen to reflect the relative importance given to the two terms in the cost function $C$.

Hypotheses about $\Phi$ will dictate learning context.

If $f$ minimizes $C$ then

$$\frac{d}{d\varepsilon} C(f + \varepsilon \overline{f}) \bigg|_{\varepsilon = 0} = 0 \quad \forall \overline{f} \in H_k$$

(1st order necessary cond.)

Compute:

$$\frac{d}{d\varepsilon} C(f + \varepsilon \overline{f}) \bigg|_{\varepsilon = 0} = \frac{1}{m} \sum_{i=1}^{m} \left< \Phi(y_i, f(x_i)) + \varepsilon \overline{f}(x_i), f + \varepsilon \overline{f}\right>_k$$

$$= \frac{1}{m} \sum_{i=1}^{m} D_2 \Phi(y_i, f(x_i)) \overline{f}(x_i) + 2 \varepsilon \left< f, \overline{f}\right>_k$$

Here $D_2 \Phi$ = partial derivative of $\Phi$ w.r.t second argument.
Set this $= 0$.

Let $\tilde{f} = K_x$. Then $\tilde{f}(x_i) = K_x(x, x_i) = K(x)$.

Thus we have the $1^{st}$ order necessary condition

$$0 = \frac{1}{m} \sum_{i=1}^{m} D_2 \Phi (y_i, f(x_i)) K_{x_i} (x)$$

$$+ 2 \lambda \langle f, K_x \rangle$$

$$= \frac{1}{m} \sum_{i=1}^{m} D_2 \Phi (y_i, f(x_i)) K_{x_i} (x)$$

$$+ 2 \lambda f(x)$$ (by reproducing property)

Hence

$$f = -\frac{1}{2 \lambda m} \sum_{i=1}^{m} D_2 \Phi (y_i, f(x_i)) K_{x_i}$$

This is known as the **representer theorem**, since it says that optimal $f$ is necessarily a linear combination of kernel functions with coefficients

$$c_i = -\frac{1}{2 \lambda m} D_2 \Phi (y_i, f(x_i))$$
\[ = - \frac{1}{2\pi m} D^2 \Phi \left( y_i, \sum_{j=1}^{m} c_j \cdot K_{x_j}(x_i) \right) \quad \text{(by representation theorem)} \]

\[ i = 1, 2, ..., m. \]

This is a system of equations for the unknown coefficients.

Suppose \( \Phi(a, b) = (a - b)^2 \). Then

\[ D^2 \Phi(a, b) = -2(a - b). \]

Hence

\[ c_i = -\frac{-2}{2\pi m} \left( y_i - \sum_{j=1}^{m} c_j \cdot K_{x_j}(x_i) \right) \quad i = 1, 2, ..., m. \]

\[ \Rightarrow \quad \sum_{i=1}^{m} c_i + \sum_{j=1}^{m} c_j \cdot K_{x_j}(x_i) = y_i, \quad i = 1, 2, ..., m \]

\[ \Rightarrow \quad (\sum_{i=1}^{m} 1 + \mathbf{K}) \mathbf{c} = \mathbf{y} \]

where we have defined \( \mathbf{K} = [K_{x_j}(x_i)] \)

\[ = [K(x_i, x_j)] \]

\[ \mathbf{c} = (c_1, ..., c_m), \quad \mathbf{y} = (y_1, ..., y_m) \quad \text{(column vectors)} \]

The matrix \( \sum_{i=1}^{m} 1 + \mathbf{K} \)

is invertible since \( \sum_{i=1}^{m} 1 > 0 \) and \( \mathbf{K} \) is positive semi-
definite, with $A$ denoting the identity matrix. Hence we can solve for unique $c$ to determine the minimizer of $C$. For general loss functions $F$, the system of equations for $c_i$, $i=1,2,\ldots,m$

$$c_i = -\frac{1}{2\lambda m} D_x F(y_i, \sum_{j=1}^{m} c_j K(x_j, x_i))$$

would be nonlinear and possibly admit multiple solutions.