Part 5 (MA 751)

Statistical machine learning and kernel methods

Primary references:

John Shawe-Taylor and Nello Cristianini, *Kernel Methods for Pattern Analysis*

Christopher Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2, 121–167 (1998).

Other references:

Aronszajn, Theory of reproducing kernels. Transactions of the American Mathematical Society, 686, 337-404, 1950.

Felipe Cucker and Steve Smale, On the mathematical foundations of learning. Bulletin of the American Mathematical Society, 2002.

Teo Evgeniou, Massimo Pontil and Tomaso Poggio, Regularization Networks and Support Vector Machines Advances in Computational Mathematics, 2000.

1. Linear functionals

Given a vector space V, we define a map $f : V \to \mathbb{R}$ from V to the real numbers to be a *functional*.

If f is *linear*, i.e., if for real a, b we have

$$f(a\mathbf{x} + b\mathbf{y}) = af(\mathbf{x}) + bf(\mathbf{y}),$$

then we say f is a *linear functional.*

If V is an inner product space (so each **v** has a length $\|\mathbf{v}\|$), we say that f is *bounded* if

$$|f(\mathbf{x})| \leq C\|\mathbf{x}\|$$

for some number $C > 0$ and all $\mathbf{x} \in X$.

2. Reproducing Kernel Hilbert spaces:

Def. 1. A $n \times n$ matrix M is symmetric if $M_{ij} = M_{ji}$ for all i, j.
A symmetric M is *positive* if all of its eigenvalues are non-negative.

Equivalently M is positive if

$$\langle \mathbf{a}, M\mathbf{a} \rangle \equiv \mathbf{a}^T M \mathbf{a} \geq 0$$

for all vectors $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$, with $\langle \cdot\,, \cdot \rangle$ the standard inner product on $\mathbb{R}^n$. Above $\mathbf{a}^T = (a_1, \ldots, a_n)$ is the transpose of $\mathbf{a}$.

Reproducing kernel Hilbert spaces

Definition 2: Let $X \subseteq \mathbb{R}^p$ be compact (i.e., a closed bounded subset). A (real) *reproducing kernel Hilbert space (RKHS)* $\mathcal{H}$ on X is a Hilbert space of functions on X (i.e., a complete collection of functions which is closed under addition and scalar mult, and for which an inner product is defined).

$\mathcal{H}$ also needs the property: for any fixed $\mathbf{x} \in X$, the evaluation functional $\mathbf{x}^* : \mathcal{H} \to \mathbb{R}$ defined by

$$\mathbf{x}^*(f) = f(\mathbf{x})$$

bounded linear functional on $\mathcal{H}$.

Reproducing kernel Hilbert spaces

Definition 3: We define a *kernel* to be a function $K : X \times X \to \mathbb{R}$ which is symmetric, i.e.,

$$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$$

for $\mathbf{x}, \mathbf{y} \in X$. We say that K is *positive* if for any fixed collection

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset X,$$

the $n \times n$ matrix

$$K = (K_{ij}) \equiv K(\mathbf{x}_i, \mathbf{x}_j)$$

is positive (i.e., non-negative).

Kernel existence

We now have the reason these are called RKHS:

Theorem 1: *Given a reproducing kernel Hilbert space $\mathcal{H}$ of functions on $X \subset \mathbb{R}^d$, there exists a unique symmetric positive kernel function $K(\mathbf{x}, \mathbf{y})$ such that for all $f \in \mathcal{H}$,*

$$f(\mathbf{x}) = \langle f(\,\cdot\,), K(\,\cdot\,, \mathbf{x}) \rangle_{\mathcal{H}}$$

(inner product above is in the variable $\cdot$; **x** is fixed).

Note this means that evaluation of f at **x** is equivalent to taking inner product of f with the fixed function $K(\,\cdot\,, \mathbf{x})$.

Kernel existence

Proof (please look at this on your own): For any fixed $\mathbf{x} \in X$, recall $\mathbf{x}^*$ is a bounded linear functional on $\mathcal{H}$. By the *Riesz Representation theorem*[1] there exists a fixed function, call it $K_{\mathbf{x}}(\,\cdot\,)$ such that for all $f \in \mathcal{H}$ (recall $\mathbf{x}$ is fixed, now f is varying)

$$f(\mathbf{x}) = \mathbf{x}^*(f) = \langle f(\,\cdot\,), K_{\mathbf{x}}(\,\cdot\,)\rangle. \tag{1}$$

(all inner products are in $\mathcal{H}$, *not* in L^2, i.e., $\langle f, g\rangle = \langle f, g\rangle_{\mathcal{H}}$).

[1] **Riesz Representation Theorem:** *If $\phi : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional on $\mathcal{H}$, there exists a unique $\mathbf{y} \in \mathcal{H}$ such that $\forall\, \mathbf{x} \in \mathcal{H},\ \phi(\mathbf{x}) = \langle \mathbf{y}, \mathbf{x}\rangle$.*

Kernel existence

That is, evaluation of f at $\mathbf{x}$ is equivalent to an inner product with the function $K_{\mathbf{x}}$.

Define $K(\mathbf{x}, \mathbf{y}) = K_{\mathbf{x}}(\mathbf{y})$. Note by (1), the functions $K_{\mathbf{x}}(\,\cdot\,)$ and $K_{\mathbf{y}}(\,\cdot\,)$ satisfy

$$\langle K_{\mathbf{x}}(\,\cdot\,), K_{\mathbf{y}}(\,\cdot\,)\rangle = K_{\mathbf{y}}(\mathbf{x}) = K_{\mathbf{x}}(\mathbf{y}),$$

so $K(\mathbf{x}, \mathbf{y})$ is symmetric.

Kernel existence

To prove $K(\mathbf{x}, \mathbf{y})$ is positive definite: let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a fixed collection. If $K_{ij} \equiv K(\mathbf{x}_i, \mathbf{x}_j)$, then if $K = (K_{ij})$ is a matrix and $\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$,

Kernel existence

$$\langle \mathbf{c}, K\mathbf{c} \rangle \equiv \mathbf{c}^T K \mathbf{c} = \sum_{i,j=1}^{n} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i,j=1}^{n} c_i c_j \langle K_{\mathbf{x}_i}(\,\cdot\,), K_{\mathbf{x}_j}(\,\cdot\,) \rangle$$

$$= \left\langle \sum_{i=1}^{n} c_i K_{\mathbf{x}_i}(\,\cdot\,), \sum_{j=1}^{n} c_j K_{\mathbf{x}_j}(\,\cdot\,) \right\rangle = \left\| \sum_{i=1}^{n} c_i K_{\mathbf{x}_i}(\,\cdot\,) \right\|_{\mathcal{H}}^2 \geq 0. \quad \square$$

Kernel existence

Definition 4: We call the above kernel $K(\mathbf{x}, \mathbf{y})$ the *reproducing kernel* of $\mathcal{H}$.

Definition 5: A *Mercer kernel* is a positive definite kernel $K(\mathbf{x}, \mathbf{y})$ which is also continuous as a function of **x** and **y** and bounded.

Def. 6: For a continuous function f on a compact set $X \subset \mathbb{R}^d$ we define

$$\|f\|_\infty = \max_{\mathbf{x} \in X} |f(\mathbf{x})|.$$

Kernel existence

Theorem 2:

(i) For every Mercer kernel $K : X \times X \to \mathbb{R}$, there exists a unique Hilbert space $\mathcal{H}$ of functions on X such that K is its reproducing kernel.

(ii) Moreover, this $\mathcal{H}$ consists of continuous functions, and for any $f \in \mathcal{H}$

$$\|f\|_\infty \leq M_K \|f\|_{\mathcal{H}},$$

where $M_K = \max_{\mathbf{x},\mathbf{y} \in X} |K(\mathbf{x}, \mathbf{y})|$.

Kernel existence

Proof (please look at this on your own): Let $K(\mathbf{x}, \mathbf{y}) : X \times X \to \mathbb{R}$ be a Mercer kernel. We will construct a reproducing kernel Hilbert space $\mathcal{H}$ with reproducing kernel K as follows.

Define

$$\mathcal{H}_0 = \text{span}\{K_{\mathbf{x}}(\,\cdot\,)\}_{\mathbf{x} \in X}$$

$$= \left\{ \sum_i c_i K_{\mathbf{x}_i}(\,\cdot\,) : \{\mathbf{x}_i\}_i \subset X \text{ is any finite subset; } c_i \in \mathbb{R} \right\}.$$

Kernel existence

Now we define inner product $\langle f, g \rangle$ for $f, g \in \mathcal{H}_0$. Assume

$$f(\,\cdot\,) = \sum_{i=1}^{l} a_i K_{\mathbf{x}_i}(\,\cdot\,), \quad g(\,\cdot\,) = \sum_{i=1}^{l} b_i K_{\mathbf{x}_i}(\,\cdot\,).$$

[Note we may assume f, g both use same set $\{\mathbf{x}_i\}$ since if not we may take a union without loss].

[Note again that here $\langle\,\cdot\,,\,\cdot\,\rangle = \langle\,\cdot\,,\,\cdot\,\rangle_{\mathcal{H}}$]

Kernel existence

Then define $\langle K_{\mathbf{x}}(\,\cdot\,), K_{\mathbf{y}}(\,\cdot\,)\rangle = K(\mathbf{x}, \mathbf{y})$

$$\langle f(\,\cdot\,), g(\,\cdot\,)\rangle = \left\langle \sum_{i=1}^{l} a_i K(\mathbf{x}_i,\,\cdot\,), \sum_{j=1}^{l} b_j K(\mathbf{x}_j,\,\cdot\,) \right\rangle$$

$$= \sum_{i,j=1}^{l} a_i a_j \langle K(\mathbf{x}_i,\,\cdot\,), K(\mathbf{x}_j,\,\cdot\,)\rangle = \sum_{i,j=1}^{l} a_i b_j K(\mathbf{x}_i, \mathbf{x}_j).$$

Kernel existence

Easy to check that with the above inner product $\mathcal{H}_0$ is an inner product space (i.e., satisfies properties $(\mathbf{a}) - (\mathbf{d})$). Now form the *completion*[2] of this space into the Hilbert space $\mathcal{H}$. Note that for $f = \sum_i a_i K_{\mathbf{x}_i}(\,\cdot\,)$ as above

---

[2]The *completion* of a non-complete inner product space space $\mathcal{H}_0$ is the (unique) smallest complete inner product (Hilbert) space $\mathcal{H}$ which contains $\mathcal{H}_0$. That is, $\mathcal{H}_0 \subset \mathcal{H}$, the inner product on $\mathcal{H}_0$ is the same as on $\mathcal{H}$, and there is no smaller complete Hilbert space which contains $\mathcal{H}_0$.

Example 1: $\mathcal{H} = \left\{ \mathbf{a} = (a_1, a_2, \dots) \,\middle|\, \sum_{i=1}^{\infty} |a_i|^2 < \infty \right\}$ with inner product $\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^{\infty} a_i b_i$ was discussed in class. The inner product space

$$\mathcal{H}_0 = \left\{ (a_1, a_2, \dots) \in \mathcal{H} \,\middle|\, \text{all but a finite number of } a_i \text{ are } 0 \right\} \subset \mathcal{H}$$

is an example of an incomplete space. $\mathcal{H}$ is its completion.

Kernel existence

$$|f(\mathbf{x})| = \langle f(\,\cdot\,), K(\mathbf{x},\,\cdot\,)\rangle \le \|f(\,\cdot\,)\|\|K(\mathbf{x},\,\cdot\,)\|$$

$$= \|f\|\sqrt{\langle K(\mathbf{x},\,\cdot\,), K(\mathbf{x},\,\cdot\,)\rangle}$$

$, a_2, \dots) \in \mathcal{H}$ | all but a finite number of a_i are $0 \subset \mathcal{H}$

is an example of an incomplete space. $\mathcal{H}$ is its completion.

Example 2: $\mathcal{H} = L^2(-\pi, \pi)$ with standard inner product for functions. We know if $f(x) \in \mathcal{H}$ then $f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx$. Define $\mathcal{H}_0$ to be all $f \in \mathcal{H}$ for which the above sum is *finite* (i.e., all but a finite number of terms are 0). Then $\mathcal{H}$ is the completion of $\mathcal{H}_0$.

$$\leq M_K \|f\|_{\mathcal{H}}.$$

[Note again here we write $\|f\| = \|f\|_{\mathcal{H}}$ by definition; similarly $\langle f, g\rangle = \langle f, g\rangle_{\mathcal{H}}$]

The above shows that the imbedding $I : \mathcal{H}_0 \to C(X)$ (the latter is the continuous functions on X) is bounded. By this we mean that I maps function f as a function in $\mathcal{H}_0$ to itself as a function in $C(X)$; in $C(X)$ the norm of f is $\|f\|_\infty \equiv \sup_{x\in X} |f(x)|$.

By bounded we mean that $\|If\|_\infty = \|f\|_\infty \leq d\|f\|_{\mathcal{H}}$ for some constant $d > 0$.

Kernel existence

Thus any Cauchy sequence in $\mathcal{H}_0$ is also Cauchy in $C(X)$, and so it follows easily that the completion $\mathcal{H}$ of $\mathcal{H}_0$ exists as a subset of $C(X)$.

That K is a reproducing kernel for $\mathcal{H}$ follows by approximation from $\mathcal{H}_0$. $\square$

3. Regularization methods for choosing f

Finding f from $Nf = \mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ is an *ill-posed problem*: the operator N^{-1} does not exist because N is not one to one.

Need to combine both:

(a) Data Nf
(b) A priori information, e.g., "f is smooth", e.g. expressing a preference for smooth over wiggly solutions seen earlier.

How to incorporate? Using *Tikhonov regularization methods.*

Regularization methods

We introduce a *regularization loss functional* $L(f)$ representing penalty for choice of an "unrealistic" f such as that in (a) above.

Assume we want to find the correct function $f_0(\mathbf{x})$, from data

$$N f_0(\mathbf{x}) = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = \mathcal{T}$$

Regularization methods

Suppose we are given $f(\mathbf{x})$ as a candidate for approximating $f_0(\mathbf{x})$ from the information in $\mathcal{T}$. We score f as a good or bad approximation based on a combination of its error on the known points $\{\mathbf{x}_i\}_{i=1}^n$, together with its "plausibility", i.e., how low the Lagrangian

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) + L(f)$$

is. Here $V(f(\mathbf{x}_i), y_i)$ is a measure of the loss whenever $f(\mathbf{x}_i)$ is far from y_i, e.g.

$$V(f(\mathbf{x}_i), y_i) = |f(\mathbf{x}_i) - y_i|^2.$$

And $L(f)$ measures the *a priori loss*, i.e., a measure of discrepancy between the prospective choice f and our prior expectation about f.

Examples: Regularization methods

Example:

$$L(f) = \|Af\|_{L^2}^2 = \int d\mathbf{x}\, |Af(\mathbf{x})|^2,$$

where $Af = -\Delta f + f$; here $\Delta f = \frac{\partial^2 f}{\partial x_1^2} + \ldots + \frac{\partial^2 f}{\partial x_p^2}$.

Note Δf and thus $L(f)$ measures the degree of non-smoothness that f has (i.e., we prefer smoother functions a priori).

Examples: Regularization methods

Example 7: Consider the case $L(f) = \|Af\|^2$ above. The norm

$$\|f\|_{\mathcal{H}} = \|Af\|_{L^2}$$

= *reproducing kernel Hilbert space norm* (at least if dimension d is small). That is, it comes from an inner product $\langle f, g\rangle = \int_X (Af)(x)(Ag)(x)dx$, and with this inner product $\mathcal{H}$ is an RKHS.

If this is the case, in general things become easier.

Examples: Regularization methods

Example 8: In the case $f = f(x),\ x \in \mathbb{R}^1$.

Suppose we choose:

$$Af = -\frac{d^2}{dx^2} f + f = \left(-\frac{d^2}{dx^2} + 1\right) f,$$

we have

$$L(f) = \|Af\|^2 = \int \left[\left(-\frac{d^2}{dx^2} + 1\right) f\right]^2 dx,$$

and $\|Af\|$ is a measure of "lack of smoothness" of f.

Examples: Regularization methods

4. More about using the Laplacian to measure smoothness (Sobolev smoothness)

Basic definitions: Recall the Laplacian operator Δ on a function f on $\mathbb{R}^p$

$$f(\mathbf{x}) = f(x_1, \ldots, x_p)$$

is defined by

$$\Delta f = \frac{\partial^2}{\partial x_1^2} f + \ldots + \frac{\partial^2}{\partial x_p^2} f.$$

Using the Laplacian for kernels

For $s > 0$ an even integer, we can define the Sobolev space H^s by:

$$H^s = \{f \in L^2(\mathbb{R}^d) : (1 - \Delta)^{s/2} f \in L^2(\mathbb{R}^p)\}$$

to be functions in $L^2(\mathbb{R}^p)$ which are still in L^2 after taking the derivative operation $(1 - \Delta)^{s/2}$, i.e., $(I - \Delta)$ repeated $s/2$ times (the operator 1 is always the identity).

For $f, g \in H^s$ define the new inner product

Using the Laplacian for kernels

$$\langle f, g \rangle_{H^s} = \langle (-\Delta + 1)^{s/2} f, (-\Delta + 1)^{s/2} g \rangle_{L^2};$$

[note $\langle h(\mathbf{x}), k(\mathbf{x}) \rangle_{L^2} = \int_X h(\mathbf{x}) k(\mathbf{x}) d\mathbf{x}$]

Can show that H^s is an RKHS with reproducing kernel

$$K(\mathbf{z}) = \mathcal{F}^{-1}\left(\frac{1}{(|\boldsymbol{\omega}|^2 + 1)^s} \right) \tag{3}$$

Using the Laplacian for kernels

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. The function $\frac{1}{(|\boldsymbol{\omega}|^2+1)^s}$ is a function on $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p) \in \mathbb{R}^p$, where $|\boldsymbol{\omega}|^2 = \omega_1^2 + \ldots + \omega_p^2$.

Using the Laplacian for kernels

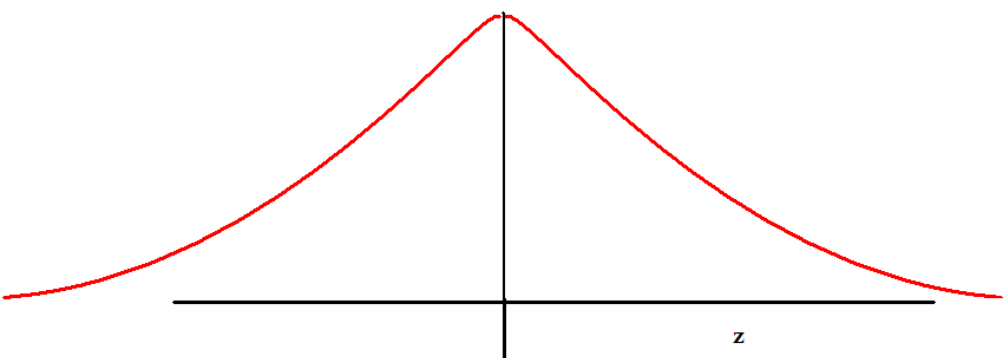


Fig 1: $K(\mathbf{z})$ in one dimension - a smooth kernel

Using the Laplacian for kernels

$K(\mathbf{z})$ is called a *radial basis function.*

Note: the kernel $K(\mathbf{x}, \mathbf{y})$ (as function of 2 variables) is defined in terms of above K by

$$K(\mathbf{x}, \mathbf{y}) = K(\mathbf{x} - \mathbf{y}).$$

Using the Laplacian for kernels

The Representer Theorem for RKHS

1. An application: using RKHS for regularization

Assume again we have an unknown function f on X, with only data

$$Nf = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) = \mathcal{T}.$$

To find the best guess $\widehat{f}$ for f, approximate it by the minimizer

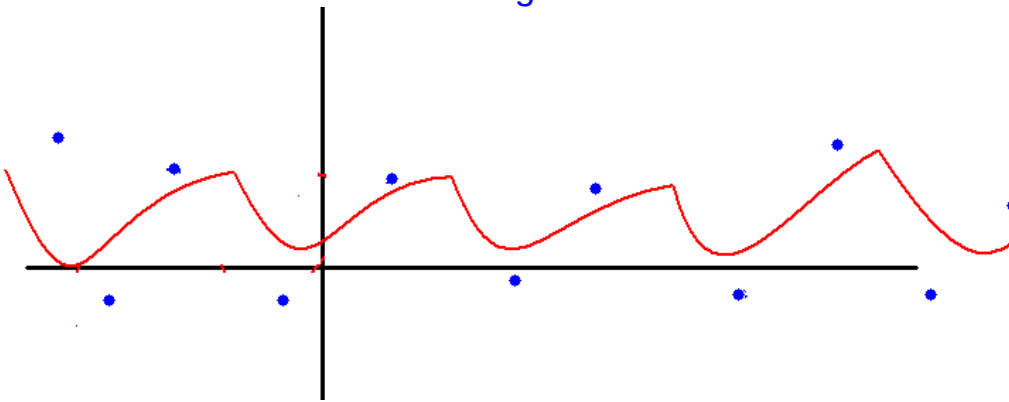
RKHS and regularization

$$\widehat{f} = \arg\min_{f \in H^s} \left\{ \frac{1}{n} \sum_{i=1}^{n} \| f(\mathbf{x}_i) - y_i \|^2 + \lambda \| f \|_{H^s}^2 \right\}. \qquad (1)$$

where λ can be some constant. Note we are finding an f which balances minimizing $\sum_{i=1}^{n} \| f(\mathbf{x}_i) - y_i \|^2$,

i.e., the data error, with minimizing $\| f \|_{H^s}^2$, i.e., maximizing the smoothness. The solution to such a problem will look like this:

RKHS and regularization



It will compromise between fitting the data (which may have error) and trying to be smooth.

The amazing thing: $\widehat{f}$ can be found explicitly using radial basis functions.

2. Solving the minimization

Now consider the optimization problem (1). We claim that we can solve it explicitly. To see this works in general for RKHS, return to the general problem.

Given an unknown $f \in \mathcal{H} =$ RKHS. Try to find the "best" approximation $\widehat{f}$ to f fitting the data

RKHS and regularization

$Nf \equiv ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$, but ALSO satisfying a priori knowledge that $\|f_0\|_{\mathcal{H}}$ is small.

Specifically, we want to find

$$\arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) + \lambda \|f\|_{\mathcal{H}}^2. \tag{2}$$

Note we can have, e.g., $V(f(\mathbf{x}_i), y_i) = (f(\mathbf{x}_i) - y_i)^2$. In that case

RKHS and regularization

$$\sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) = \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2.$$

Consider the general case (2), with arbitrary error measure V.
We have the

RKHS and regularization

Representer Theorem: *A solution of the Tikhonov optimization problem* (1) *can be written*

$$\widehat{f}(x) = \sum_{i=1}^{n} a_i K(\mathbf{x}, \mathbf{x}_i), \tag{3}$$

where K *is the reproducing kernel of the RKHS* $\mathcal{H}$*.*

Important theorem: says we only need to find a set of n numbers a_i to optimize the infinite dimensional problem (1) above.

Proof: Use calculus of variations. If a minimizer f_1 exists, then for all $g \in \mathcal{H}$, assuming that the derivatives with respect to ϵ exist:

Representer theorem proof

$$0 = \frac{d}{d\epsilon}\frac{1}{n}\sum_{i=1}^{n} V((f_1 + \epsilon g)(\mathbf{x}_i), y_i) + \lambda \|f_1 + \epsilon g\|_{\mathcal{H}}^2 \Big|_{\epsilon=0}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial V}{\partial f_1(\mathbf{x}_i)}(f_1(\mathbf{x}_i), y_i) \cdot g(\mathbf{x}_i)$$

$$+ \lambda \frac{d}{d\epsilon}\left\{\langle f_1, f_1\rangle + 2\epsilon\langle f_1, g\rangle + \epsilon^2\langle g, g\rangle\right\}$$

Representer theorem proof

$$= \frac{1}{n}\sum_{i=1}^{n} V_1(f_1(\mathbf{x}_i), y_i) \cdot g(\mathbf{x}_i) + 2\lambda\langle f_1, g\rangle,$$

where $V_1(a,b) = \frac{\partial}{\partial a}V(a,b)$ and all inner products are in $\mathcal{H}$.

Since the above is true for all $g \in \mathcal{H}$, it follows that if we let $g = K_{\mathbf{x}}$ we get:

Representer theorem proof

$$0 = \frac{1}{n}\sum_{i=1}^{n} V_1(f_1(\mathbf{x}_i), y_i) K_{\mathbf{x}}(\mathbf{x}_i) + 2\lambda \langle f_1, K_{\mathbf{x}} \rangle$$

$$= \frac{1}{n}\sum_{i=1}^{n} V_1(f_1(\mathbf{x}_i), y_i) K_{\mathbf{x}}(\mathbf{x}_i) + 2\lambda f_1(\mathbf{x}),$$

or

$$f_1(\mathbf{x}) = \frac{1}{2\lambda n}\sum_{i=1}^{n} V_1(f_1(\mathbf{x}_i), y_i) K(\mathbf{x}, \mathbf{x}_i).$$

Representer theorem proof

Thus if a minimizer $\widehat{f} = f_1$ exists for (1), it can be written in the form (3) as claimed, with

$$a_i = \frac{1}{2\lambda n} V_1(f_1(\mathbf{x}_i), y_i).$$

Note that this does not solve the problem, since the a_i are expressed in terms of the solution itself. But it does reduce the possibilities for what a solution looks like.