Least-Squares in (∞-dimensional) Hilbert Space

Let's look back at our solution to the ridge regression problem when $\boldsymbol{A}$ is underdetermined (more columns than rows, $M > N$). We solve

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \delta\|\boldsymbol{x}\|_2^2,$$

which we can also rewrite as

$$\underset{\boldsymbol{x}}{\text{minimize}}\ \ \sum_{m=1}^{M} \left|y_m - \langle \boldsymbol{x}, \boldsymbol{a}_m\rangle\right|^2 + \delta\|\boldsymbol{x}\|_2^2,$$

where the $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M \in \mathbb{R}^M$ are the rows of $\boldsymbol{A}$ (after we transpose them):

$$\boldsymbol{A} = \begin{bmatrix} -\boldsymbol{a}_1^{\mathrm{T}}- \\ -\boldsymbol{a}_2^{\mathrm{T}}- \\ \vdots \\ -\boldsymbol{a}_M^{\mathrm{T}}- \end{bmatrix}.$$

We have seen that in this case the solution is

$$\hat{\boldsymbol{x}} = \left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A} + \delta\mathbf{I}\right)^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{y} = \boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \delta\mathbf{I}\right)^{-1}\boldsymbol{y}$$

We explicitly point out two facts about the solution:

1. $\hat{\boldsymbol{x}}$ is in the *row space* (the linear span of the rows) of $\boldsymbol{A}$

$$\hat{\boldsymbol{x}} = \boldsymbol{A}^{\mathrm{T}}\hat{\boldsymbol{\alpha}} = \sum_{m=1}^{M}\hat{\alpha}_m\,\boldsymbol{a}_m.$$

2. The coefficients $\hat{\boldsymbol{\alpha}}$ are computed by solving the symmetric positive definite system of equations

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + \delta\mathbf{I})^{-1}\boldsymbol{y}, \quad \boldsymbol{K} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}.$$

The $M \times M$ matrix $\boldsymbol{K}$ is formed by taking all the inner products between the different rows of $\boldsymbol{A}$:

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_1 & \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_2 & \cdots & \boldsymbol{a}_1^{\mathrm{T}}\boldsymbol{a}_M \\ \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_1 & \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_2 & \cdots & \boldsymbol{a}_2^{\mathrm{T}}\boldsymbol{a}_M \\ \vdots & & \ddots & \vdots \\ \boldsymbol{a}_M^{\mathrm{T}}\boldsymbol{a}_1 & \cdots & & \boldsymbol{a}_M^{\mathrm{T}}\boldsymbol{a}_M \end{bmatrix}.$$

We will see below that these two facts extend to the analogous problem in an ∞-dimensional Hilbert space ... this is convenient in that it allows us to solve ∞-dimensional optimization problems with a finite amount of computational effort.

Let $\mathcal{S}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_S$ — for the purposes of discussion, we will think of the elements of $\mathcal{S}$ as containing functions that map $\mathbb{R}^D \rightarrow \mathbb{R}$. Let $\boldsymbol{f}$ be a function in $\mathcal{S}$ that we are trying to estimate from evaluations against a series of continuous linear functionals represented by $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M \in \mathcal{S}$:

$$\begin{aligned} y_1 &= \langle \boldsymbol{f}, \boldsymbol{a}_1 \rangle_S \ + \text{noise} \\ y_2 &= \langle \boldsymbol{f}, \boldsymbol{a}_2 \rangle_S \ + \text{noise} \\ &\vdots \\ y_M &= \langle \boldsymbol{f}, \boldsymbol{a}_M \rangle_S \ + \text{noise} \end{aligned}$$

This is analagous to observing $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \text{noise}$ in the finite dimensional case (where $\boldsymbol{x} \in \mathbb{R}^N$), where each entry in $\boldsymbol{y}$ is being modeled as an inner product between a row in $\boldsymbol{A}$ and $\boldsymbol{x}$. Here, it's sort of like we are observing $\boldsymbol{y} \in \mathbb{R}^M$ through a "matrix" that has M rows, but each row is a function instead of a vector in $\mathbb{R}^N$.

Suppose now that we estimate $\boldsymbol{f}$ by solving the following least-squares problem in $\mathcal{S}$:

$$\underset{\boldsymbol{f}\in\mathcal{S}}{\text{minimize}} \quad \sum_{m=1}^{M} \left|y_m - \langle \boldsymbol{f}, \boldsymbol{a}_m\rangle_S\right|^2 + \delta\|\boldsymbol{f}\|_S^2. \tag{1}$$

This is an optimization program in an infinite dimensional Hilbert space. Even if we discretize the search space using an orthobasis (or any basis), there will be an infinite number of expansion coefficients to solve for. However, the following result (which is almost an immediate consequence of our work on linear approximation from a few weeks ago) shows us that we can at least specify the solution by solving an $M \times M$ system of equations.

Representer Theorem (least-squares version):
The solution to (1) ia given by

$$\hat{\boldsymbol{f}} = \sum_{m=1}^{M} \hat{\alpha}_m \, \boldsymbol{a}_m,$$

where

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{K}+\delta\mathbf{I})^{-1}\boldsymbol{y}, \quad \boldsymbol{K} = \begin{bmatrix} \langle \boldsymbol{a}_1, \boldsymbol{a}_1\rangle_S & \langle \boldsymbol{a}_2, \boldsymbol{a}_1\rangle_S & \cdots & \langle \boldsymbol{a}_M, \boldsymbol{a}_1\rangle_S \\ \langle \boldsymbol{a}_1, \boldsymbol{a}_2\rangle_S & \langle \boldsymbol{a}_2, \boldsymbol{a}_2\rangle_S & \cdots & \langle \boldsymbol{a}_M, \boldsymbol{a}_2\rangle_S \\ \vdots & & \ddots & \vdots \\ \langle \boldsymbol{a}_1, \boldsymbol{a}_M\rangle_S & & \cdots & \langle \boldsymbol{a}_M, \boldsymbol{a}_M\rangle_S \end{bmatrix}.$$

Proof. We use the notation

$$L(\boldsymbol{f}) = \sum_{m=1}^{M} \left|y_m - \langle \boldsymbol{f}, \boldsymbol{a}_m\rangle_S\right|^2,$$

and so we are trying to solve

$$\underset{\boldsymbol{f}}{\text{minimize}}\, L(\boldsymbol{f}) + \delta\|\boldsymbol{f}\|_2^2.$$

Let $\mathcal{A} = \operatorname{span}(\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M\})$ be the subspace spanned by the $\boldsymbol{a}_m$ (and since the $\boldsymbol{a}_m$ are linearly independent, they are a basis for $\mathcal{A}$). For any candidate function $\boldsymbol{g} \in \mathcal{S}$, we can write

$$\boldsymbol{g} = \boldsymbol{g}_A + \boldsymbol{g}_\perp,$$

where $\boldsymbol{g}_A$ is the closest point in $\mathcal{A}$ to $\boldsymbol{g}$ and $\boldsymbol{g}_\perp = \boldsymbol{g} - \boldsymbol{g}_A$ is orthogonal to every vector in $\mathcal{A}$; in particular

$$\langle \boldsymbol{g}_\perp, \boldsymbol{a}_m \rangle_S = 0, \quad m = 1, \ldots, M.$$

Then

$$\begin{aligned} L(\boldsymbol{g}) &= \sum_{m=1}^{M} |y_m - \langle \boldsymbol{g}_A + \boldsymbol{g}_\perp, \boldsymbol{a}_m \rangle_S|^2 \\ &= \sum_{m=1}^{M} |y_m - \langle \boldsymbol{g}_A, \boldsymbol{a}_m \rangle_S|^2 \quad (\text{since } \langle \boldsymbol{g}_\perp, \boldsymbol{a}_m \rangle_S = 0) \\ &= L(\boldsymbol{g}_A), \end{aligned}$$

and

$$\begin{aligned} \|\boldsymbol{g}\|_S^2 &= \|\boldsymbol{g}_A\|_S^2 + \|\boldsymbol{g}_\perp\|_S^2 \quad (\text{Pythagorean thm.}) \\ &\geq \|\boldsymbol{g}_A\|_S^2. \end{aligned}$$

Thus

$$L(\boldsymbol{g}) + \delta\|\boldsymbol{g}\|_S^2 \;\geq\; L(\boldsymbol{g}_A) + \delta\|\boldsymbol{g}_A\|_S^2.$$

So for every $\boldsymbol{g} \in \mathcal{S}$, there is a corresponding member of $\mathcal{A}$ that makes the functional we are trying to minimize at least as small. Thus at least one solution to (1) must be in $\mathcal{A}$, and we can write

$$\hat{\boldsymbol{f}} = \sum_{m=1}^{M} \hat{\alpha}_m \, \boldsymbol{a}_m, \tag{2}$$

for some $\hat{\alpha}_1, \ldots, \hat{\alpha}_M$. Note that if the $\boldsymbol{a}_m$ are not all linearly independent, then there might be multiple $\hat{\boldsymbol{\alpha}}$ that give the same $\hat{\boldsymbol{f}}$, but this is not a problem; we only need to find one of them. No matter what, we still have that $\hat{\boldsymbol{f}}$ is the **unique** solution to (1), there just might be multiple ways to write that solution.

We pause here to specifically point out something that will make everything a little easier below. If $\boldsymbol{v}$ and $\boldsymbol{u}$ are elements in $\boldsymbol{\mathcal{A}}$ with

$$\boldsymbol{u} = \sum_{m=1}^{M} c_m \boldsymbol{a}_m, \quad \text{and} \quad \boldsymbol{v} = \sum_{m=1}^{M} d_m \boldsymbol{a}_m,$$

then

$$\begin{aligned}\langle \boldsymbol{u}, \boldsymbol{v} \rangle_S &= \left\langle \sum_{m=1}^{M} c_m \boldsymbol{a}_m, \sum_{\ell=1}^{M} d_\ell \boldsymbol{a}_\ell \right\rangle_S \\ &= \sum_{m=1}^{M} \sum_{\ell=1}^{M} c_m d_\ell \langle \boldsymbol{a}_m, \boldsymbol{a}_\ell \rangle_S \\ &= \boldsymbol{d}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{c},\end{aligned}$$

where $\boldsymbol{K}$ is the same matrix given in the statement of the theorem (the Gram matrix for the $\{\boldsymbol{a}_m\}$).

To find the best $\boldsymbol{\alpha}$ in (2), we solve

$$\begin{gathered}\underset{\boldsymbol{\alpha}\in\mathbb{R}^M}{\text{minimize}} \sum_{m=1}^{M} \left| y_m - \sum_{\ell=1}^{M} \alpha_\ell \langle \boldsymbol{a}_\ell, \boldsymbol{a}_m \rangle \right|^2 + \delta \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha} \\ \downarrow \\ \underset{\boldsymbol{\alpha}\in\mathbb{R}^M}{\text{minimize}} \sum_{m=1}^{M} |y_m - (\boldsymbol{K}\boldsymbol{\alpha})_m|^2 + \delta \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha} \\ \downarrow \\ \underset{\boldsymbol{\alpha}\in\mathbb{R}^M}{\text{minimize}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}\|_2^2 + \delta \boldsymbol{\alpha}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{\alpha}.\end{gathered}$$

Taking the gradient of the functional above and setting it equal to zero[1] tells us that $\hat{\boldsymbol{\alpha}}$ must obey

$$\boldsymbol{K}^{\mathrm{T}}(\boldsymbol{K}\hat{\boldsymbol{\alpha}} - \boldsymbol{y}) + \delta \boldsymbol{K}\hat{\boldsymbol{\alpha}} = \boldsymbol{0}.$$

The matrix $\boldsymbol{K}$ is symmetric (so $\boldsymbol{K}^{\mathrm{T}} = \boldsymbol{K}$) and invertible, so the condition above is equivalent to

$$(\boldsymbol{K} + \delta\mathbf{I})\hat{\boldsymbol{\alpha}} = \boldsymbol{y}.$$

Since the matrix $\boldsymbol{K} + \delta\mathbf{I}$ is always invertible, the theorem is established.

■

[1] The gradient (first derivative) being zero is always a necessary condition for minimizer of a functional with multiple arguments. In this case, it happens to be sufficient as well since the Hessian matrix (second derivative) is positive definite, meaning the functional is convex.

Supervised learning in a RKHS (Kernel regression)

We start by recalling our central regression problem: we are given pairs of data points $(\boldsymbol{t}_m, y_m)$, $m = 1, \ldots, M$ with $\boldsymbol{t}_m \in \mathbb{R}^D$ and $y_m \in \mathbb{R}$, and we want to find a mapping $\boldsymbol{f} : \mathbb{R}^D \to \mathbb{R}$ such that

$$f(\boldsymbol{t}_m) \approx y_m, \quad m = 1, \ldots, M. \tag{3}$$

Making this problem well-posed relies critically on having some kind of *model* for $\boldsymbol{f}$; we have seen already how to set up and solve it when we model $\boldsymbol{f}$ as being in an N-dimensional space spanned by a basis of our choosing. In this section, we consider a more general model: $\boldsymbol{f}$ is a member of a reproducing kernel Hilbert space (RKHS) $\mathcal{S}$ with kernel $k(\cdot, \cdot)$.

We have seen that in an RKHS, each sample (or point evaluation) of a function $\boldsymbol{f}$ can be written as an inner product against another known, fixed function in the same RKHS. We can re-write (3) as

$$\langle \boldsymbol{f}, \boldsymbol{k}_{t_m} \rangle \approx y_m, \quad m = 1, \ldots, M.$$

The function $\boldsymbol{k}_{t_m} \in \mathcal{S}$ is of course different for every different location $\boldsymbol{t}_m \in \mathbb{R}^D$. As before, we define the **kernel** for the RKHS as the ensemble of all of these functions; we use the notation

$$\boldsymbol{k}(\cdot, \boldsymbol{t}_m) = \boldsymbol{k}_{t_m}(\cdot),$$

where $k(\cdot, \cdot) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is the aforementioned kernel for $\mathcal{S}$. Given $\{(\boldsymbol{t}_m, y_m)\}$, we estimate the mapping $\boldsymbol{f}$ by solving the infinite-dimensional least-squares problem:

$$\underset{\boldsymbol{f} \in \mathcal{S}}{\text{minimize}} \ \sum_{m=1}^{M} |y_m - f(\boldsymbol{t}_m)|^2 + \delta \|\boldsymbol{f}\|_S^2$$

which is the same as

$$\underset{\boldsymbol{f}\in\mathcal{S}}{\text{minimize}} \;\; \sum_{m=1}^{M} |y_m - \langle \boldsymbol{f}, \boldsymbol{k}_{t_m}\rangle|^2 + \delta\|\boldsymbol{f}\|_S^2.$$

We now know exactly how to solve this problem. First, we form the $M \times M$ matrix $\boldsymbol{K}$ with entries that are inner products between all of the different sampling functions

$$K_{i,j} = \langle \boldsymbol{k}_{t_j}, \boldsymbol{k}_{t_i}\rangle = \langle \boldsymbol{k}(\cdot, \boldsymbol{t}_j), \boldsymbol{k}_{t_i}\rangle = k(\boldsymbol{t}_i, \boldsymbol{t}_j).$$

That is, the $\boldsymbol{K}$ matrix is just the kernel function $k(\cdot,\cdot)$ evaluated at all the pairs of input points. We then solve

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + \delta\mathbf{I})^{-1}\boldsymbol{y},$$

to get our estimate

$$\hat{\boldsymbol{f}} = \sum_{m=1}^{M} \hat{\alpha}_m \boldsymbol{k}_{t_m}.$$

The $\hat{\boldsymbol{\alpha}}$ gives us an implicit representation for $\hat{\boldsymbol{f}}$. When it comes time to use this function, that is evaluate it at some point $\boldsymbol{\tau}$ not in the data set $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_M$, we have

$$\hat{f}(\boldsymbol{\tau}) = \sum_{m=1}^{M} \hat{\alpha}_m k_{t_m}(\boldsymbol{\tau}) = \sum_{m=1}^{M} \hat{\alpha}_m k(\boldsymbol{\tau}, \boldsymbol{t}_m).$$

This procedure for fitting $\hat{\boldsymbol{f}}$ to data points $\{\boldsymbol{t}_m, y_m\}_{m=1}^{M}$ is called **kernel regression**.

Let's reflect on the expressions above with two observations, one old, one new:

1. We solved an infinite-dimensional optimization problem using finite-dimensional linear algebra (thank you again, orthogonality principle!). The main computation is solving an $M \times M$ systems of equations — that is, the computation scales with the amount of data we have rather than the dimension the space we are in (which is good, since that dimension is ∞).
2. If we are given the kernel $k(\cdot, \cdot)$, even the inner products are completely abstracted away ... both the forming of the $\boldsymbol{K}$ matrix and the evaluation of $\hat{\boldsymbol{f}}$ after the estimate has been formed rely only on evaluating $k(\cdot, \cdot)$ at different points.

Observation 2 is especially interesting. In some sense, we don't even need to know the Hilbert space $\mathcal{S}$ or its inner product $\langle \cdot, \cdot \rangle_S$ to solve this problem ... we only need to know the kernel. We now have another modeling tool for fitting functions, rather than model them as coming from certain finite-dimensional subspaces (and then representing them using a basis for that subspace), we can model them as coming from an (infinite-dimesional) RKHS with a certain kernel. Different kernels lead to different models (and hence different estimates), but the computations to solve the least-squares problem always follow the same steps above.

It is natural at this point to ask which $k(\cdot, \cdot)$ can be the kernel for an RKHS. Indeed, we will ask this question and then answer it in excruciating detail. But first, let's discuss some of the similarities and differences in the basis and RKHS function models.