Least-Squares in (∞-dimensional) Hilbert Space

We have seen that introducing a (finite dimensional) basis allows us to turn the regression problem into a finite-dimensional linear algebra problem. In this section, we will see that even without a basis, we can solve least-squares problems in ∞-dimensional Hilbert space by recasting them as finite-dimensional linear algebra problems.

To start, 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

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

which we can also rewrite as

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \ \sum_{m=1}^{M} \left|y_m - \boldsymbol{a}_m^{\mathrm{T}}\boldsymbol{x}\right|^2 + \delta\|\boldsymbol{x}\|_2^2,$$

where the $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M \in \mathbb{R}^N$ 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 $\mathbb{R}^D \to \mathbb{R}$, but the math we use below applies to any abstract Hilbert space. Let $\boldsymbol{f}$ be a function in $\mathcal{S}$ that we are trying to estimate from observations of a series of inner products against fixed $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M \in \mathcal{S}$:

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

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 in $\mathcal{S}$ instead of a vector in $\mathbb{R}^N$.

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

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

This is an optimization program in a (possibly) 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 a little while 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

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

Let $\mathcal{A} = \text{Span}\left(\{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_M\}\right)$ be the subspace spanned by the $\boldsymbol{a}_m$. 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

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

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 $\mathcal{A}$ 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 — it is similar to the Gram matrix for the $\{\boldsymbol{a}_m\}$, though there is no

guarantee here that the $\{\boldsymbol{a}_m\}$ are linearly independent, and so $\boldsymbol{K}$ might not be invertible.

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

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

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

$$\boldsymbol{K}^{\mathrm{T}}(\boldsymbol{K}\hat{\boldsymbol{\alpha}} - \boldsymbol{y}) + \delta \boldsymbol{K}\hat{\boldsymbol{\alpha}} = \boldsymbol{0}. \tag{3}$$

The matrix $\boldsymbol{K}$ is symmetric ($\boldsymbol{K}^{\mathrm{T}} = \boldsymbol{K}$), and so any $\hat{\boldsymbol{\alpha}}$ that obeys

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

will also obey (3). Since the matrix $\boldsymbol{K} + \delta\mathbf{I}$ is always invertible, we can take $\hat{\boldsymbol{\alpha}} = (\boldsymbol{K} + \delta\mathbf{I})^{-1}\boldsymbol{y}$ and 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 semi-definite, meaning the functional is convex.