3 Variational splines and representer theorems

The leading idea is that splines can be defined as solutions of variational problems subject to some convex linear measurement constraints. While such splines are defined over the continuum, the pleasing outcome of the theory is that they live in a finite-dimensional subspace that is tied to the underlying regularization operator.

Given a suitable RKHS $\mathcal{H}$, we like to view a generalized spline as a function $f \in \mathcal{H}$ that is uniquely characterized through the values $z_m = \langle \nu_m, f \rangle$ of a finite number M of linear functionals $\nu_1, \ldots, \nu_M \in \mathcal{H}'$. The traditional setting corresponds to the choice of the sampling functionals $\nu_m = \delta(\cdot - \boldsymbol{x}_m)$, which translates into the (non-uniform) interpolation conditions $f(\boldsymbol{x}_m) = y_m$, $m = 1, \ldots, M$. Since the specification of the value of these linear functionals is obviously not enough to determine f unambiguously, the unicity of the spline is achieved by minimizing the corresponding *spline energy* $\|\mathrm{L}f\|_{L_2}^2$ where L is an admissible regularization operator.

The notion of spline is also suitable for dealing with approximate or perturbed measurements $y_m = \langle \nu_m, f \rangle + \epsilon_m$ where ϵ_m is some unknown disturbance term that is typically assumed to be random (noise) and identically distributed for each component. The best fitting spline is then determined by minimizing the least-squares functional

$$J_{\mathrm{LS}}(f|\mathbf{y}, \lambda) = \sum_{m=1}^{M} \big(y_m - \langle \nu_m, f \rangle\big)^2 + \lambda \|\mathrm{L}f\|_{L_2(\mathbb{R}^d)}^2 \tag{96}$$

where L is the regularization operator and $\lambda \in \mathbb{R}^+$ is a tradeoff parameter that controls the closeness of the fit. We note that the exact fit with $y_m = z_m = \langle \nu_m, f \rangle$ is achieved by letting $\lambda \to 0$.

3.1 Regularization functional induced by an inner product

Let us start with the easier scenario where the optimization is performed over a Hilbert space $\mathcal{H}$ and the regularization functional is the quadratic norm induced by its inner product. To set up the problem, it is convenient to recall that the effect of a measurement functional $\nu_m \in \mathcal{H}'$ (the continuous dual of $\mathcal{H}$) has an equivalent representation as

$$\langle \nu_m, f \rangle = \langle \nu_m^*, f \rangle_{\mathcal{H}}$$

where $\nu_m^* = \mathrm{R}\{\nu_m\} \in \mathcal{H}$ is the (unique) Riesz conjugate of ν_m (see Theorem 4).

Our first approach for deriving the spline solution is to recast (96) as the minimization of a generic quadratic form covered by Theorem 35 in Appendix A. To that end, we rewrite $J_{\mathrm{LS}}(f|\mathbf{y},\lambda)$ as

$$J_{\mathrm{LS}}(f|\mathbf{y},\lambda)=\sum_{m=1}^{M}\left(y_m-\langle\nu_m^*,f\rangle_{\mathcal{H}}\right)^2+\lambda\|f\|_{\mathcal{H}}^2=\tfrac{1}{2}a(f,f)-v(f)+C_0$$

with

$$C_0=\sum_{m=1}^{M}y_m^2,\qquad v(f)=2\sum_{m=1}^{M}y_m\langle\nu_m^*,f\rangle_{\mathcal{H}}$$
$$a(f_1,f_2)=2\lambda\langle f_1,f_2\rangle_{\mathcal{H}}+2\sum_{m=1}^{M}\langle\nu_m^*,f_1\rangle_{\mathcal{H}}\langle\nu_m^*,f_2\rangle_{\mathcal{H}},$$

where $v=\sum_{m=1}^{M}y_m\nu_m^*\in\mathcal{H}'$ is a continuous linear functional on $\mathcal{H}$ and $a:\mathcal{H}\times\mathcal{H}\to\mathbb{R}$ a continuous symmetric bilinear form on $\mathcal{H}$. We also note that the bilinear form is coercive when $\lambda>0$. Theorem 35 then tells us that the minimizer of this functional, $f_0=\arg\min_{f\in\mathcal{H}}J(f|\mathbf{y},\lambda)$, is unique and such that $v(f)=a(f_0,f)$, which translates into

$$\langle\sum_{m=1}^{M}y_m\nu_m^*,f\rangle_{\mathcal{H}}=\langle\lambda f_0,f\rangle_{\mathcal{H}}+\langle\sum_{m=1}^{M}\langle\nu_m^*,f_0\rangle_{\mathcal{H}}\nu_m^*,f\rangle_{\mathcal{H}}$$

for all $f\in\mathcal{H}$. This is equivalent to

$$\lambda f_0=\sum_{m=1}^{M}\left(y_m-\langle\nu_m^*,f_0\rangle_{\mathcal{H}}\right)\nu_m^*,\tag{97}$$

which implies that $f_0\in\mathrm{span}\{\nu_m^*\}_{m=1}^{M}$, or, stated explicitly,

$$f_0=\sum_{m=1}^{M}a_m\nu_m^*$$

for some suitable weights $a_1,\ldots,a_M\in\mathbb{R}$. Upon substitution of this latter expansion in (97), we find that the optimal coefficient vector $\mathbf{a}=(a_1,\ldots,a_M)$ is the solution of the linear system of equations

$$(\mathbf{G}+\lambda\mathbf{I}_M)\mathbf{a}=\mathbf{y}\tag{98}$$

where $\mathbf{y} = (y_1, \ldots, y_M)$ is the data vector and $\mathbf{G} = \langle \boldsymbol{\nu}^*, (\boldsymbol{\nu}^*)^T \rangle_{\mathcal{H}}$ the $M \times M$ Gram (or correlation) matrix whose entries are given by

$$[\mathbf{G}]_{m,n} = \langle \nu_m^*, \nu_n^* \rangle_{\mathcal{H}} = \langle \nu_m, \nu_n \rangle_{\mathcal{H}'} = \langle \nu_m, \nu_n^* \rangle.$$

For further reference, we summarize the result in Proposition 15 while we also provide the concrete representation associated with a RKHS.

Proposition 15 (Regularized least-squares approximation)**.** *Let us consider the following.*

- $\mathcal{H}$ *is a Hilbert space associated with the inner product* $\langle f, g \rangle_{\mathcal{H}}$*;*
- $\nu_1, \ldots, \nu_M \in \mathcal{H}'$ *is a finite set of linear measurement functionals on* $\mathcal{H}$ *with corresponding Riesz conjugates* $\nu_1^*, \ldots, \nu_M^* \in \mathcal{H}$ *such that* $\langle \nu_m, f \rangle = \langle \nu_m^*, f \rangle_{\mathcal{H}}$ *for all* $f \in \mathcal{H}$*;*
- $\mathbf{G} \in \mathbb{R}^{M \times M}$ *is a symmetric positive-definite matrix with entries* $[\mathbf{G}]_{m,n} = \langle \nu_m, \nu_n^* \rangle = \langle \nu_m^*, \nu_n^* \rangle_{\mathcal{H}}$ *and* $\mathbf{I}_M$ *the identity matrix of size* M*;*
- $\mathbf{y} = (y_1, \ldots, y_M) \in \mathbb{R}^M$ *is some arbitrary data vector and* $\lambda \in \mathbb{R}^+$ *an adjustable regularization parameter.*

Then, the abstract regularized least-squares reconstruction problem

$$\arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} \left| y_m - \langle \nu_m, f \rangle \right|^2 + \lambda \langle f, f \rangle_{\mathcal{H}}^2 \right) \tag{99}$$

has a unique solution that is given by

$$f = \sum_{m=1}^{M} a_m \nu_m^* \quad \text{with} \quad \mathbf{a} = (a_1, \ldots, a_M) = (\mathbf{G} + \lambda \mathbf{I}_M)^{-1} \mathbf{y}.$$

In particular, when $\mathcal{H}$ *is a RKHS over* $\mathbb{R}^d$ *with reproducing kernel* $r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{y})$*, the minimizer* f *admits the functional representation*

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} a_m \nu_m^*(\boldsymbol{x}), \qquad \nu_m^*(\boldsymbol{x}) = \int_{\mathbb{R}^d} r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{y}) \nu_m(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}$$

while the entries of $\mathbf{G}$ *can be evaluated as*

$$[\mathbf{G}]_{m,n} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \nu_m(\boldsymbol{x}) r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{y}) \nu_n(\boldsymbol{y}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{y}. \tag{100}$$

Proof. There are many ways of establishing the result. The simplest is to define $\mathcal{V} = \text{span}\{\nu_m^*\}_{m=1}^M$ and to use the same of orthogonality argument as in the proof of Theorem 18 to show that the minimizer $f_0 \in \text{span}\{\nu_m^*\}_{m=1}^M$, while $\mathbf{G}$ is Gram matrix of the underlying basis. Once the parametric form of the solution is established, one simply verifies that $\boldsymbol{\nu}(f) = \mathbf{G}\mathbf{a}$ and $\|f\|_{\mathcal{H}}^2 = \mathbf{a}^T\mathbf{G}\mathbf{a}$ so that the optimal optimal expansion coefficients are found by minimizing the quadradic cost

$$J(\mathbf{a}|\mathbf{y}, \lambda) = \|\mathbf{y} - \mathbf{G}\mathbf{a}\|_2^2 + \lambda \mathbf{a}^T\mathbf{G}\mathbf{a}.$$

This is achieved by evaluating the partial derivative of J with respect to $\mathbf{a}$,

$$\frac{\partial J(\mathbf{a}|\mathbf{y}, \lambda)}{\partial \mathbf{a}} = 2\mathbf{G}(\mathbf{y} - \mathbf{G}\mathbf{a}) + 2\lambda\mathbf{G}\mathbf{a} = 2\mathbf{G}\big(\mathbf{y} - (\mathbf{G} + \lambda\mathbf{I}_M)\mathbf{a}\big).$$

This gradient clearly vanishes if we set $\mathbf{y} = (\mathbf{G} + \lambda\mathbf{I}_M)\mathbf{a}$, which yields the desired solution. In fact, the latter condition is also necessary for optimality, due to the unicity of the solution (see Hilbert's projection Theorem 34). □

We observe that the approach can be taken to the limit with $\lambda \to 0$ whenever $\mathbf{G}$ is invertible. Moreover, we can readily determine the "spline energy" of the solution as

$$\begin{aligned}\|f\|_{\mathcal{H}}^2 &= \langle f, f\rangle_{\mathcal{H}} \\ &= \sum_{m=1}^M a_m \sum_{n=1}^M a_n \langle \nu_m^*, \nu_n^*\rangle_{\mathcal{H}} = \mathbf{a}^T\mathbf{G}\mathbf{a},\end{aligned}$$

which is a quadratic form associated with the Gram matrix $\mathbf{G}$.

Proposition 15 provides us with a simple linear algorithm for the resolution of regularized least-squares problems. We shall now generalize the result and show that the parametric form of the solution is preserved for a much broader class of optimization problems.

Theorem 18 (Abstract representer theorem)**.** *Let $\mathcal{H}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and Riesz map $\mathrm{R} : \mathcal{H}' \to \mathcal{H}$. Let $\boldsymbol{\nu} : \mathcal{H} \to \mathbb{R}^M : f \mapsto \boldsymbol{\nu}(f) = (\langle\nu_1, f\rangle, \ldots, \langle\nu_M, f\rangle)$ with $\nu_m \in \mathcal{H}'$ be a continuous linear measurement operator and $\mathcal{C}$ be a closed convex subset of $\mathbb{R}^M$ such that its preimage in $\mathcal{H}$, $\mathcal{U} = \boldsymbol{\nu}^{-1}(\mathcal{C}) = \{f \in \mathcal{H} : \boldsymbol{\nu}(f) \in \mathcal{C}\}$, is nonempty (feasibility hypothesis). Then, the problem*

$$\arg\min_{f\in\mathcal{H}} \|f\|_{\mathcal{H}}^2 \quad s.t. \quad \boldsymbol{\nu}(f) \in \mathcal{C} \tag{101}$$

has a unique solution of the form

$$f_0 = \sum_{m=1}^{M} a_m \nu_m^* \tag{102}$$

with $\nu_m^* = \mathrm{R}\{\nu_m\} \in \mathcal{H}$ *and suitable weights* $a_m \in \mathbb{R}$ *for* $m = 1, \ldots, M$.

Proof. The unicity of the solution follows from Hilbert's projection theorem (Theorem 34). In the present context, we are projecting the origin $f = 0$ onto the convex set $\mathcal{U}$, which is nonempty because of the feasibility hypothesis. The enabling property is that convexity (resp., closedness) is preserved through linear (resp., continuous) transformations so that the preimage $\mathcal{U}$ of the closed convex set $\mathcal{C}$ is guaranteed to be closed and convex as well.

Next, we invoke Riesz' representation theorem (Theorem 4) and rewrite $\langle \nu_m, f \rangle = \langle \nu_m^*, f \rangle_{\mathcal{H}}$, for $m = 1, \ldots, M$, where $\nu_m^* = \mathrm{R}\{\nu_m\} \in \mathcal{H}$ is the conjugate of $\nu_m \in \mathcal{H}'$. Defining $\mathcal{V} = \mathrm{span}\{\nu_m^*\}_{m=1}^M$, we then specify $\mathcal{V}^\perp = \{f \in \mathcal{H} : \langle f, \nu_m^* \rangle_{\mathcal{H}} = 0, m = 1, \ldots, M\}$ as the orthogonal complement of $\mathcal{V}$ in $\mathcal{H}$. The key is then to observe that $\mathcal{V}^\perp$ coincides with the null space of the measurement operator $\boldsymbol{\nu}$. Since $\mathcal{H} = \mathcal{V} \oplus \mathcal{V}^\perp$, every $f \in \mathcal{H}$ has a unique decomposition as $f = u + u^\perp$ with $u \in \mathcal{V}$ and $u^\perp \in \mathcal{V}^\perp$. The solution f_0 can therefore be written as $f_0 = u_0 + u_0^\perp$ with $\boldsymbol{\nu}(f_0) = \boldsymbol{\nu}(u_0)$ and $\boldsymbol{\nu}(u_0^\perp) = 0$, which implies that u_0 also lies in $\mathcal{V}$. Since f_0 is the minimal-norm solution, we have that

$$\begin{aligned} \|f_0\|_{\mathcal{H}}^2 \leq \|u_0\|_{\mathcal{H}}^2 \quad &\Rightarrow \|u_0 + u_0^\perp\|_{\mathcal{H}}^2 = \|u_0\|_{\mathcal{H}}^2 + \|u_0^\perp\|_{\mathcal{H}}^2 \leq \|u_0\|_{\mathcal{H}}^2 \\ &\Rightarrow \|u_0^\perp\|_{\mathcal{H}} = 0 \Leftrightarrow u_0^\perp = 0 \end{aligned}$$

Thus, $f_0 = u_0$ implying that $f_0 \in \mathrm{span}\{\nu_m^*\}_{m=1}^M$, which is equivalent to (102).

□

Let us now briefly show that the result in Theorem 18 is also applicable to our initial quadratic minimization problem (96). Given the data vector $\mathbf{y} = (y_1, \ldots, y_M)$, the underlying loss function is the quadratic error

$$F_2(\mathbf{z}, \mathbf{y}) = \|\mathbf{y} - \mathbf{z}\|^2 \quad \text{with} \quad \mathbf{z} = \boldsymbol{\nu}(f),$$

which is a continuous convex function of $\mathbf{z} \in \mathbb{R}^M$ and, by composition, of $f \in \mathcal{H}$ since the measurement map $f \mapsto \mathbf{z} = \boldsymbol{\nu}(f)$ is linear and bounded. By considering the level sets of $F_2(\cdot, \mathbf{y})$, one obtains a series of embedded closed convex sets

$$\mathcal{C}_{\mathbf{y},\sigma} = \{\mathbf{z} \in \mathbb{R}^N : \|\mathbf{y} - \mathbf{z}\|_2^2 \leq \sigma^2\}$$

that are parametrized by $\sigma^2 \geq 0$. The application of Theorem 18 ensures the existence of a unique solution $f_{(\mathbf{y},\sigma)}$ such that $\boldsymbol{\nu}(f_{(\mathbf{y},\sigma)}) \in \mathcal{C}_{\mathbf{y},\sigma}$ and $\|\mathrm{L}f\|_2^2$ is minimum. Moreover, since the projection of a function onto a closed convex set is necessarily located on the frontier of that set, we have $\|\mathbf{y} - \boldsymbol{\nu}(f_{(\mathbf{y},\sigma)})\|^2 = \sigma^2$. On the other hand, since the solution of our initial problem $f_\lambda = \arg\min J(f|\mathbf{y}, \lambda)$ with λ fixed is unique, there exists a corresponding "optimal" $\sigma_\lambda = \sigma(\lambda) \geq 0$ such that $\|\mathbf{y} - \boldsymbol{\nu}(f_\lambda)\|_2^2 = \sigma_\lambda^2$. It then follows that the two problems are equivalent if we set $\sigma = \sigma_\lambda$ with $f_{(\mathbf{y},\sigma_\lambda)} = f_\lambda$.

The extreme scenario is $\sigma = 0$, which yields an interpolating solution such that $\boldsymbol{\nu}(f_{(\mathbf{y},0)}) = \mathbf{y}$. The latter also corresponds to the solution of (96) as $\lambda \to 0$.

Now, the truly powerful aspect of Theorem 18 is that the above reasoning remains applicable for non-quadratic loss functionals, subject to some mild convexity constraints, as further discussed in Sections 3.1.2 and 3.2.3.

3.1.1 Smoothing splines and ridge regression

As already announced in the introduction, the classical scenario of spline interpolation corresponds to the choice of a series of ideal sampling functionals $\nu_m = \delta(\cdot - \boldsymbol{x}_m)$ with $\langle \nu_m, f \rangle = f(\boldsymbol{x}_m), m = 1, \ldots, M$. Since the location of the $\boldsymbol{x}_m \in \mathbb{R}^d$ can be arbitrary, the sampling is called *non-uniform*.

In the event where the data is noisy, one considers a relaxed form of interpolation also known as a *smoothing* spline. Mathematically, this is formulated as a regularized least-squares recovery problem

$$f_{\mathrm{LS}} = \arg\min_{f \in \mathcal{H}} \left(\sum_{m=1}^{M} \big(y_m - f(\boldsymbol{x}_m)\big)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right)$$

where the choice of the quadratic data term—that is, the loss function $F_2\big(\boldsymbol{\nu}(f), \mathbf{y}\big) = \|\mathbf{z} - \mathbf{y}\|^2$ with $z_m = f(\boldsymbol{x}_m)$—is primarily motivated by computational convenience.

Once more, the general form of the solution is given by (102) in Theorem 18. In the present case, this simplifies to

$$f_{\mathrm{LS}}(\boldsymbol{x}) = \sum_{m=1}^{M} a_m \mathrm{R}\{\delta(\cdot - \boldsymbol{x}_m)\}(\boldsymbol{x}) = \sum_{m=1}^{M} a_m r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_m) \tag{103}$$

where we have used the identity $r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_m) = \mathrm{R}\{\delta(\cdot - \boldsymbol{x}_m)\}(\boldsymbol{x})$, which is a restatement of the definition of the reproducing kernel (see Property 6 in

Proposition 2). Based on (98), we also obtain the expression of the expansion coefficients

$$\mathbf{a}_{\text{LS}} = (a_1, \ldots, a_M) = (\mathbf{R} + \lambda \mathbf{I}_M)^{-1}\mathbf{y}, \tag{104}$$

where the generalized Gram matrix $\mathbf{G}$ in (98) reduces to the symmetric inner-product matrix $\mathbf{R} \in \mathbb{R}^{M\times M}$ with entry

$$[\mathbf{R}]_{m,n} = \langle r_{\mathcal{H}}(\cdot, \boldsymbol{x}_m), r_{\mathcal{H}}(\cdot, \boldsymbol{x}_n)\rangle_{\mathcal{H}} = r_{\mathcal{H}}(\boldsymbol{x}_m, \boldsymbol{x}_n). \tag{105}$$

Moreover, the use of Property 2 in Proposition 2 yields

$$\|f_{\text{LS}}\|^2_{\mathcal{H}} = \mathbf{a}_{\text{LS}}^T\, \mathbf{R}\, \mathbf{a}_{\text{LS}},$$

which provides the spline energy of the solution.

In statistical regression, one usually assumes that the samples of the signal are corrupted by AWG noise of variance σ^2. The minimum mean square error (MMSE) reconstruction of the signal is then given by the same formula as (103) with the reproducing kernel $r_{\mathcal{H}}(\cdot, \boldsymbol{x}_m)$ being substituted by the covariance function of the signal $r_f(\cdot, \boldsymbol{x}_m) = \mathbb{E}\{f(\cdot)f(\boldsymbol{x}_m)\}$ and $\lambda = \sigma^2$ (see Section 4.6). The corresponding estimation method is called *ridge regression* since the qualitative effect of (104) is to add a ridge (i.e., a constant diagonal term of height σ^2) to the covariance matrix $\mathbf{R}$ of the signal. The same type of estimator is also used in geostatistics where is known as kriging.

The good news with the smoothing spline problem is that $(\mathbf{R} + \lambda\mathbf{I})$ with $\lambda \geq 0$ is always invertible (including for the limit case $\lambda \to 0$) because $\mathbf{R}$ is symmetric positive-definite, as a consequence of the strict positive definiteness of the reproducing kernel (see Definition 2). The geometric interpretation of this results is that the functions $r_{\mathcal{H}}(\cdot, \boldsymbol{x}_m)$ are linearly independent; hence they provide a bona fide basis for representing the solution of spline-related optimization problems.

For further reference, we specify our "optimal" spline interpolant as f_{int}: it is the minimum-norm member of $\mathcal{H}$ that satisfies the interpolation constraints $f_{\text{int}}(\boldsymbol{x}_m) = y_m$ for $m = 1, \ldots, M$. f_{int} lives in the M-dimensional subspace specified by (103) and is uniquely characterized by its expansion coefficients

$$\mathbf{a}_{\text{int}} = \mathbf{R}^{-1}\mathbf{y}.$$

In fact, we can vary $\mathbf{y}$ to generate the whole spline space that is spanned by $\{r_{\mathcal{H}}(\cdot, \boldsymbol{x}_m)\}_{m=1}^M$, while we have the property that $\|f_{\text{int}}\|^2_{\mathcal{H}} = \mathbf{a}_{\text{int}}^T\mathbf{R}\mathbf{a}_{\text{int}} = \mathbf{y}^T\mathbf{R}^{-1}\mathbf{y} = \mathbf{y}^T\mathbf{a}_{\text{int}}$.

Description	Definition	Convex	Coercive
Square loss	$\|\boldsymbol{z}-\mathbf{y}\|_2^2=\sum_{m=1}^{M}(z_m-y_m)^2$	Yes	Yes
Absolute loss	$\|\boldsymbol{z}-\mathbf{y}\|_1=\sum_{m=1}^{M}\|z_m-y_m\|$	Yes	Yes
Hinge loss	$\sum_{m=1}^{M}\max(1-z_m y_m)$	Yes	Yes

3.1.2 Representer theorem for statistics and machine learning

The next step in the progression is machine learning because it typically involves loss functions that are more sophisticated than the least-square criterion in (96). The measurement functionals, on the other hand, are still kept simple with $\nu_m=\delta(\cdot-\boldsymbol{x}_m)$ and $\langle\delta(\cdot-\boldsymbol{x}_m),f\rangle=f(\boldsymbol{x}_m)$.

In essence, the problem is to find a function $f:\mathbb{R}^d\to\mathbb{R}$ such that $f(\boldsymbol{x}_m)\approx y_m$ (with a training set of size M) where the proximity (or loss) between the samples $\mathbf{z}(f)=\big(f(\boldsymbol{x}_1),\ldots,f(\boldsymbol{x}_M)\big)$ and $\mathbf{y}=(y_1,\ldots,y_M)$ is measured by some loss function $\mathbb{R}^M\times\mathbb{R}^M\to\mathbb{R}$.

The two key properties that help us ensure the unicity of the solution are convexity and coercivity.

Definition 15 (Convex function)**.** *A multivariate function $g:\mathbb{R}^M\to\mathbb{R}$ is convex if*

$$g\big(\tau\mathbf{z}_1+(1-\tau)\mathbf{z}_2\big)\le\tau g\big(\mathbf{z}_1\big)+(1-\tau)g\big(\mathbf{z}_2,\mathbf{y}\big)$$

for all $\mathbf{z}_1,\mathbf{z}_2\in\mathbb{R}^M$ and all $\tau\in[0,1]$. It is strictly convex if the order relation holds with a strict inequality for $\mathbf{z}_1\neq\mathbf{z}_2$ and $\tau\in(0,1)$.

An important property of finite-dimensional convex functions is that they are continuous inside their domain (see [**?**, Corollary 2.3, p. 12]).

Definition 16 (Coercive function)**.** *A multivariate function $g:\mathbb{R}^M\to\mathbb{R}$ is said to be coercive if $\lim_{\|\boldsymbol{z}\|\to\infty}g(\mathbf{z})=\infty$.*

With a slight abuse of language, we shall say that the loss function $F:\mathbb{R}^M\times\mathbb{R}^M\to\mathbb{R}$ is convex (resp., coercive) if $F(\cdot,\mathbf{y}):\mathbb{R}^M\to\mathbb{R}$ is convex (resp., coercive) for any fixed $\mathbf{y}\in\mathbb{R}^d$. For instance, the condition is automatically met when the cost functional can be written as $F(\mathbf{z},\mathbf{y})=g(\mathbf{z}-\mathbf{y})$ where g is a convex (resp., coercice) function.

In the statement of our optimization problem, the first argument of the loss function is not defined on $\mathbb{R}^M$, but rather on some Hilbert space $\mathcal{H}$ via the composition of some finite-dimensional linear map $f \mapsto \mathbf{z}(f)$. The main point for our argumentation is that this composition preserves convexity and continuity under the assumption that $\nu_m \in \mathcal{H}'$. The final ingredient is the regularization functional $f \mapsto \|f\|_{\mathcal{H}}^2$, which is generally quadratic and associated with a given RKHS $\mathcal{H}$.

The fundamental result for machine learning is the celebrated Representer Theorem which is often used to justify the use of kernel methods such as support vector machines (SVM), radial basis functions (RBF) and kernel PCA. This is a very powerful result whose proof is actually much simpler as one would expect.

Theorem 19 (Representer theorem for machine learning)**.** *Let $\mathcal{H}$ be a RKHS space with reproducing kernel $r_{\mathcal{H}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We consider the samples $\mathbf{z}(f) = \big(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_M)\big)$ with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathbb{R}^d$ of a function $f \in \mathcal{H}$ and corresponding data values $\mathbf{y} = (y_1, \ldots, y_M) \in \mathbb{R}^M$. Then, the solution of the generic minimisation problem*

$$\arg\min_{f \in \mathcal{H}} \big(F\big(\mathbf{z}(f), \mathbf{y}\big) + \lambda \|f\|_{\mathcal{H}}^2\big)\,, \tag{106}$$

where the cost function F is strictly convex, is unique and of the form

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} a_m r_{\mathcal{H}}(\boldsymbol{x}, \boldsymbol{x}_m) \tag{107}$$

with suitable weights $a_1, \ldots, a_M \in \mathbb{R}$.

Proof. Since the map $f \mapsto \mathbf{z}(f)$ is linear and continuous, the functional $f \mapsto F\big(\mathbf{z}(f), \mathbf{y}\big)$ is strictly convex and continuous on $\mathcal{H}$ for any fixed $\mathbf{y} \in \mathbb{R}^N$ (see Appendix B for the definition of the relevant properties in the functional setting). Likewise, the regularization term $\lambda\|f\|_{\mathcal{H}}^2$ is strictly convex, continuous and trivially coercive on $\mathcal{H}$ in the sense of Definition 39. It follows that the functional in (106) is strictly convex, continuous (and, a fortiori, lower semicontinuous) and coercive over $\mathcal{H}$, which ensures that the problem has a unique minimizer f_0 (by Proposition 27 in Appendix B). This solution achieves some "optimal" sample values $\mathbf{z}(f_0) = \mathbf{y}_0 = (y_{0,1}, \ldots, y_{0,M}) \in \mathbb{R}^M$, which fixes the data term (or loss functional) in (106) to $F\big(\mathbf{z}(f_0), \mathbf{y}\big) = F(\mathbf{y}_0, \mathbf{y})$. By imposing the condition $\mathbf{z}(f) = \mathbf{y}_0$, we freeze the data term, which allows us to reformulate the minimization of (106) as a classical spline

interpolation problem

$$\min_{f\in\mathcal{H}} \|f\|_{\mathcal{H}}^2 \quad \text{s.t.} \quad (f(\boldsymbol{x}_m) = y_{0,m})_{m=1}^M.$$

From the analysis in Section 3.1.1, we know that this problem admits a unique solution $f_{\text{int}} = f_0$ given by (103) with $\mathbf{a}_{\text{int}} = \mathbf{R}^{-1}\mathbf{y}_0$, which is consistent with (107). □

In view of the discussion following Theorem 11, we can also formulate the constrained version of the optimization problem by considering the convex set

$$\mathcal{C}_{\mathbf{y},\sigma} = \{\mathbf{z} \in \mathbb{R}^N : F\big(\mathbf{z}, \mathbf{y}\big) \leq \sigma^2\}$$

where $\sigma \in \mathbb{R}^+$ is a suitable bound on the loss. Moreover, the solution f_0 of the constrained problem is such that $F\big(\mathbf{z}(f_0), \mathbf{y}\big) = \sigma^2$ since the minimum is generally achieved on the frontier of the convex set (by Hilbert's projection theorem). Thus, we may divide (106) by λ and interpret $(1/\lambda)$ as the Lagrange multiplier associated with the norm-minimization problem with an equality constraint on the loss. At any rate, the main point is that, whatever the formulation—unconstrained or constrained with a bound on the loss or a bound on the regularization—the parametric form (107) of the solution remains the same. It is universal in the sense that it is independent of the choice of the loss function provided that the latter is convex. The caveat, of course, is that the determination of the optimal a_m generally requires the deployment of an iterative solver; typically, some type of steepest descent algorithm.

3.2 Non-coercive regularization functionals

The construction of generalized splines is also possible for the type of non-coercive regularization operator $\mathrm{L} : \mathcal{H}_{\mathrm{L}} \to L_2(\mathbb{R}^d)$ considered in Section 2.7. The operators that are suitable for this purpose are the ones that are spline-admissible in the sense of Definition 13. They are characterized by a non-trivial, finite-dimensional null space $\mathcal{N}_{\mathrm{L}}$ that admits some biorthogonal system $\{\phi_n, p_n\}_{n=1}^{N_0}$ with the property that

$$p = \mathrm{Proj}_{\mathcal{N}_{\mathrm{L}}}\{p\} = \sum_{n=1}^{N_0} \langle \phi_n, p\rangle p_n = \boldsymbol{p}^T \boldsymbol{\phi}(p)$$

for all $p \in \mathcal{N}_{\mathrm{L}} = \mathrm{span}\{p_n\}_{n=1}^{N_0}$. The corresponding native space $\mathcal{H}_{\mathrm{L}}$ is the Hilbert space $\mathcal{H}_{\mathrm{L}}$ equipped with the composite norm

$$\|f\|_{\mathcal{H}_{\mathrm{L}}} = \sqrt{\|\mathrm{L}f\|_{L_2}^2 + \|\boldsymbol{\phi}(f)\|_2^2},$$

where the second component $\|\boldsymbol{\phi}(f)\|_2 = \|\boldsymbol{\phi}(p)\|_2$ with $p = \mathrm{Proj}_{\mathcal{N}_\mathrm{L}}\{f\}$ is required to remove the ambiguity for the elements of $\mathcal{H}_\mathrm{L}$ that are in the null space of L. Since $\|\mathrm{L}f\|_{L_2}$ is only a semi-norm, it is harder to ensure unicity which calls for a more involved analysis.

Supplementary material:

3.2.1 Generalized boundary value problem

The simplest solution for avoiding any potential unicity problem is to fix the null-space component of the solution, which is achieved by imposing suitable boundary conditions. This results in a generalized boundary value problem that falls within the general Hilbert-space framework of Theorem 18. While the proposed reformulation deviates from our initial optimization problem, we shall see that the framework is rich enough to encompass the unconstrained scenarios that are covered by the representer Theorems 19 and 20.

Corollary 4. *Let us consider the following:*

- $(\mathrm{L}, \boldsymbol{\phi})$ *is an admissible pair and* $\boldsymbol{p} = (p_1, \ldots, p_{N_0})$ *a corresponding biorthogonal basis of* $\mathcal{N}_\mathrm{L}$ *such that* $\langle \phi_m, p_n \rangle = \delta_{m-n}$;
- $\boldsymbol{\nu} : f \mapsto \boldsymbol{\nu}(f) = (\langle \nu_1, f \rangle, \ldots, \langle \nu_M, f \rangle)$ *is a bounded linear measurement operator from* $\mathcal{H}_\mathrm{L} \to \mathbb{R}^M$;
- $\mathcal{C}$ *is a closed convex subset of* $\mathbb{R}^M$ *and* $\boldsymbol{c} = (c_0, \ldots, c_{N_0}) \in \mathbb{R}^{N_0}$ *a constant vector such that the set* $\mathcal{U}_{\boldsymbol{c}} = \{f \in \mathcal{H}_\mathrm{L} : \boldsymbol{\nu}(f) \in \mathcal{C}$ *and* $\boldsymbol{\phi}(f) = \boldsymbol{c}\}$ *is nonempty (feasibility hypothesis).*

Then, the solution of the minimization problem

$$\begin{aligned} \arg\min_{f \in \mathcal{H}_\mathrm{L}} \|\mathrm{L}f\|^2_{L_2(\mathbb{R}^d)} \quad s.t.\ \boldsymbol{\nu}(f) \in \mathcal{C}, \\ \langle \phi_1, f \rangle = c_1 \\ \vdots \\ \langle \phi_{N_0}, f \rangle = c_{N_0} \end{aligned}$$

is unique and of the form

$$f_0 = \sum_{m=1}^{M} a_m \mathrm{A}_{\boldsymbol{\phi}}\{\nu_m\} + \sum_{n=1}^{N_0} c_n p_n \tag{108}$$

where the $a_m \in \mathbb{R}$ *are some suitable weights and where* $\mathrm{A}_{\boldsymbol{\phi}} = (\mathrm{L}_{\boldsymbol{\phi}}^{-1}\mathrm{L}_{\boldsymbol{\phi}}^{-1*}) : \mathcal{H}'_{\mathrm{L}} \to \mathcal{H}_{\mathrm{L}}$ *is the linear operator whose kernel is specified by* (59) *in Theorem 11.*

Proof. Since $\mathcal{H}_{\mathrm{L}}$ is the direct sum of $\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$ and $\mathcal{N}_{\mathrm{L}}$, every element $f \in \mathcal{H}_{\mathrm{L}}$ has a unique decomposition as $f = g + q$ with $g \in \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$ and $q \in \mathcal{N}_{\mathrm{L}}$. Conversely, for any $g \in \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$ and $q \in \mathcal{N}_{\mathrm{L}}$, we have that $f = g+q \in \mathcal{H}_{\mathrm{L}}$ with the property that $\|\mathrm{L}f\|_{L_2(\mathbb{R}^d)} = \|\mathrm{L}g\|_{L_2(\mathbb{R}^d)} = \|g\|_{\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}}$. Moreover, since $\boldsymbol{\phi}(g) = \mathbf{0}$ for all $g \in \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$ and thanks to the biorthogonality of $(p_n, \phi_n)_{n=1}^{N_0}$, we can enforce the boundary conditions $\boldsymbol{\phi}(f) = \mathbf{c}$ by taking the null-space component as

$$q_0 = \sum_{n=1}^{N_0} c_n p_n.$$

This allows us to rewrite the optimization problem as

$$g_0 = \arg \min_{g \in \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}} \|g\|^2_{\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}} \text{ s.t. } \boldsymbol{\nu}(g) \in \mathcal{C}_{q_0}$$

where $\mathcal{C}_{q_0} = \{\mathbf{z} : \boldsymbol{z} + \boldsymbol{\nu}(q_0) \in \mathcal{C}\}$ is a closed convex set of $\mathbb{R}^M$ that is the translated version of $\mathcal{C}$ by $\boldsymbol{\nu}(q_0)$. The solution g_0 is then derived from Theorem 18, while the corresponding reproducing kernel and factorization of the Riesz map $\mathrm{A}_{\boldsymbol{\phi}} : \mathcal{H}'_{\mathrm{L},\boldsymbol{\phi}} \to \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$ is obtained from Theorem 11. The solution of our initial problem is $f_0 = g_0 + q_0$. □

In principle, the generic form of the expansion (102) is also transferable to the unconstrained scenario with the caveat that we first need to make sure that such a solution exists.

In the absence of boundary conditions, the limiting factor is the lack of coercivity of the regularization functional $\|\mathrm{L}f\|^2_{L_2}$. This forces us to impose constraints on the operator L to ensure the existence and unicity of the solution.

3.2.2 Proper regularization of an inverse problem

The regularization has obviously no effect on the null-space component of the signal. Accordingly, we must ensure that the measurements $\boldsymbol{\nu}(f)$ are rich enough to characterize this component unambiguously. This property is uncapsuled in the following definition.

Definition 17 (Proper regularization operator)**.** *The operator* $\mathrm{L} : \mathcal{H}_{\mathrm{L}} \to L_2(\mathbb{R}^d)$ *with finite-dimensional null space* $\mathcal{N}_{\mathrm{L}} = \mathrm{span}\{p_n\}_{n=1}^{N_0} \subseteq \mathcal{H}_{\mathrm{L}}$ *is a proper regularization operator for the measurement operator* $\boldsymbol{\nu} : f \mapsto \boldsymbol{\nu}(f) = (\langle \nu_1, f\rangle, \ldots, \langle \nu_M, f\rangle)$ *if the following technical conditions are met:*

1. L *is spline-admissible in the sense of Definition 13*
2. $\nu_1, \ldots, \nu_M \in \mathcal{H}'_{\mathrm{L}}$
3. *For all* $q \in \mathcal{N}_{\mathrm{L}}$, $\|\boldsymbol{\nu}(q)\|_2 \geq 0$ *with equality if and only if* $q = 0$.

Condition 2 is merely a restatement of the fact that the linear measurement operator $\boldsymbol{\nu}$ must well defined (i.e., bounded) on the full native space $\mathcal{H}_{\mathrm{L}}$. The critical requirement is Condition 3, which has the following implications.

Proposition 16 (Criteria for a proper regularization)**.** *Let* L *and* $\boldsymbol{\nu}$ *be such that the two first conditions in Definition 17 are satisfied. Then, the third condition for a proper regularization can be restated in any of the following equivalent forms.*

1. *For any* $q_1, q_2 \in \mathcal{N}_{\mathrm{L}}$, $\displaystyle\sum_{m=1}^{M} |\langle \boldsymbol{\nu}_n, q_1 \rangle - \langle \boldsymbol{\nu}_n, q_2 \rangle|^2 = 0 \Leftrightarrow q_1 = q_2$
2. *For any basis* $\{p_n\}_{n=1}^{N_0}$ *of* $\mathcal{N}_{\mathrm{L}}$*, the singular values of the* $M \times N_0$ *matrix* $\mathbf{P} = [\boldsymbol{\nu}(p_1) \ \cdots \ \boldsymbol{\nu}(p_{N_0})]$ *are strictly positive and bounded.*
3. *For any biorthogonal basis* $\{\phi_n, p_n\}_{n=1}^{N_0}$ *of* $\mathcal{N}_{\mathrm{L}}$*, there exists a constant* $c > 0$ *such that, for all* $q \in \mathcal{N}_{\mathrm{L}}$*,*

$$\|q\|^2 = \sum_{n=1}^{N_0} |\langle \phi_n, q \rangle|^2 \leq \frac{1}{c} \sum_{m=1}^{M} |\langle \nu_n, q \rangle|^2. \tag{109}$$

4. *Within* $\{\nu_m\}_{m=1}^{M}$ *with* $M \geq N_0$*, there exists at least one subset of* N_0 *functionals that are* linearly independent *on* $\mathcal{N}_{\mathrm{L}}$*. Hence, by assuming that the* ν_m *are ordered such that these* N_0 *functionals come first, we can truncate the sum over* m *in Statements 1 and 3 to the first* N_0 *terms only.*
5. *For any basis* $\{p_n\}_{n=1}^{N_0}$ *of* $\mathcal{N}_{\mathrm{L}}$*, there exists a set of biothogonal functionals* $\phi_1, \cdots, \phi_{N_0} \in \mathrm{span}\{\nu_n\}_{n=1}^{N_0}$ *where the underlying* ν_n *satisfy the linear independence property identified in 4.*

Proof. Item 1 is obvious: Since $\boldsymbol{\nu}$ is linear, the statement $\|\boldsymbol{\nu}(q)\|_2 = 0 \Leftrightarrow q = 0$ is equivalent to saying that any $q \in \mathcal{N}_{\mathrm{L}}$ is uniquely determined by its measurements $\mathbf{b} = \boldsymbol{\nu}(q)$. If we now expand q as $q = \sum_{n=1}^{N_0} c_n p_n = \boldsymbol{p}^T \mathbf{c}$, this results in the overdetermined system of equations $\mathbf{P}\mathbf{c} = \mathbf{b}$ where $\mathbf{P}$ is the

cross-product matrix of size $M \times N_0$ specified in Item 2. It is well known that such a linear system admits a unique solution

$$\mathbf{c} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{b} \tag{110}$$

(which is also the least squares one) if and only if the normal matrix $(\mathbf{P}^T\mathbf{P})$ is invertible. The latter is equivalent to the singular values of $\mathbf{P}$ being bounded away from 0, as stated in Item 2. As for Item 3, we convert (110) in the following inequality

$$\|q\| = \|\mathbf{c}\|_2 \leq \frac{\sigma_{\max}(\mathbf{P})}{\sigma^2_{\min}(\mathbf{P})}\|\boldsymbol{\nu}(q)\|_2$$

where $0 < \sigma_{\min}(\mathbf{P})$ and $\sigma_{\max}(\mathbf{P}) < \infty$ are the minimum and maximum singular values of $\mathbf{P}$, respectively (the boundedness of $\sigma_{\max}(\mathbf{P})$ simply follows from the assumption $\nu_m \in \mathcal{H}'_{\mathrm{L}}$).

Let $\mathbf{P} = [\mathbf{p}_1 \ \cdots \ \mathbf{p}_M]^T$ with $\mathbf{p}_m \in \mathbb{R}^{N_0}$. The geometric implication of Statement 2 is that the vectors $\{\mathbf{p}_m\}_{m=1}^M$ span the Euclidean space $\mathbb{R}^{N_0}$. Accordingly, when $M > N_0$, we can recursively drop dependent row vectors from $\mathbf{P}$ such as to end up with a subset of size N_0 that forms a basis of $\mathbb{R}^{N_0}$. In other words, we can ensure that $\mathbb{R}^{N_0} = \mathrm{span}\{\mathbf{p}_n\}_{n=1}^{N_0}$ modulo some proper reordering of the vectors. This implies that the reduced matrix $\mathbf{P}_0 = [\mathbf{p}_1 \ \cdots \ \mathbf{p}_{N_0}]^T$ of size N_0 satisfies the stability condition in Item 2, which takes care of Statement 4. Moreover, this ensures that the reduced cross-correlation matrix $\mathbf{P}_0$ is invertible.

As for the last statement, it translates into the selection of the dual space $\mathcal{N}'_{\mathrm{L}} = \mathrm{span}\{\nu_n\}_{n=1}^{N_0} \subseteq \mathcal{H}'_{\mathrm{L}}$, which admits a unique basis $\{\phi_n\}_{n=1}^{N_0}$ that is biorthogonal to $\{p_n\}$. The constructive procedure is the dual of the one in Proposition 9: Specifically, $\boldsymbol{\phi} = \mathbf{P}_0^{-1}\boldsymbol{\nu}_0$ where $\boldsymbol{\nu}_0 = (\nu_1, \cdots, \nu_{N_0})$ denotes our reduced vector of measurement functionals that are linearly independent.

□

3.2.3 Representer theorem for linear inverse problems

The other related issue is that the mere convexity of the cost function $F : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ is no longer sufficient to ensure unicity. To regain control over the null-space component of the signal, we need to add the coercivity requirement.

Theorem 20 (Representer theorem for linear inverse problems)**.** *Let us consider the following.*

- $\boldsymbol{\nu} : f \mapsto \boldsymbol{\nu}(f) = (\langle \nu_1, f \rangle, \ldots, \langle \nu_M, f \rangle)$ *is a bounded linear operator* $\mathcal{H}_{\mathrm{L}} \to \mathbb{R}^M$ *that extracts* M *measurements from the signal* f*;*
- $\mathrm{L} : \mathcal{H}_{\mathrm{L}} \to L_2(\mathbb{R}^d)$ *is a proper regularization operator with respect to* $\boldsymbol{\nu}$ *in the sense of Definition 17;*
- $\{p_n\}_{n=1}^{N_0}$ *is a basis of the null space of the regularization operator;*
- $F : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$ *is a continuous loss function that is strictly convex and coercive in its first argument;*
- $\mathbf{y} \in \mathbb{R}^M$ *is a given data vector and* $\lambda \in \mathbb{R}^+$ *an adjustable regularization parameter.*

Then, the solution of the generic minimization problem

$$\arg\min_{f \in \mathcal{H}_{\mathrm{L}}} J(f|\mathbf{y}, \lambda) \quad \textit{with} \quad J(f|\mathbf{y}, \lambda) = F(\boldsymbol{\nu}(f), \mathbf{y}) + \lambda \|\mathrm{L}f\|^2_{L_2(\mathbb{R}^d)} \tag{111}$$

is unique and of the form

$$f_{(\lambda)} = \sum_{m=1}^{M} a_m \varphi_m + \sum_{n=1}^{N_0} b_n p_n \tag{112}$$

with

$$\varphi_m = \mathrm{A}\{\nu_m\} = \int_{\mathbb{R}^d} G_{\mathrm{L}^*\mathrm{L}}(\cdot, \boldsymbol{y}) \nu_m(\boldsymbol{y}) \mathrm{d}\boldsymbol{y} \tag{113}$$

where $\mathbf{a} = (a_1, \ldots, a_M) \in \mathbb{R}^M$ *and* $\mathbf{b} = (b_1, \ldots, b_{N_0}) \in \mathbb{R}^{N_0}$ *are suitable coefficient vectors and where* $G_{\mathrm{L}^*\mathrm{L}}$ *is a symmetric Green's function of* $\mathrm{L}^*\mathrm{L}$*, as specified in Theorem 11. Moreover, the leading term in* (112) *is "orthogonal" to the second in the sense that* $\langle \mathbf{a}, \boldsymbol{\nu}(p_n) \rangle = 0$ *for* $n = 1, \ldots, N_0$*.*

Proof. The proof is similar to the one of Theorem 19, except that we now also need to establish the coercivity and strict convexity of $J : \mathcal{H}_{\mathrm{L}} \to \mathbb{R}$. To that end, we use the property that any element $f \in \mathcal{H}_{\mathrm{L}}$ has a unique decomposition as $f = \mathrm{L}_{\boldsymbol{\phi}}^{-1} w + q$ with $w = \mathrm{L}f$, $q = \mathrm{Proj}_{\mathcal{N}_{\mathrm{L}}}\{f\}$, and $\|f\|^2 = \|w\|^2_{L_2} + \|q\|^2$ with $\|q\| = \|\boldsymbol{\phi}(q)\|_2$ (see Theorem 12).

(i) Coercivity of J: Imposing $\|f\| \to \infty$ forces at least one of the norm components $\|w\|_{L_2}$ or $\|q\|$ to grow to ∞. Now, the lower bound (109) and the coercivity of F imply that $F\big(\boldsymbol{\nu}(\mathrm{L}_{\boldsymbol{\phi}}^{-1} w) + \boldsymbol{\nu}(q), \mathbf{y}\big) \to \infty$ as $\|q\| \to \infty$, while

the coercivity of the regularization term with respect to the w component is obvious. This implies that $J(f|\mathbf{y},\lambda) \to \infty$ as $\|f\| \to \infty$.

(ii) Strict-convexity of J: Let us pick some $w \in L_2(\mathbb{R}^d)$ and set $\tilde{f} = \mathrm{L}_{\boldsymbol{\phi}}^{-1} w$. Then, for any $q_1, q_2 \in \mathcal{N}_{\mathrm{L}}$, we define $f_1 = \tilde{f} + q_1$, $f_2 = \tilde{f} + q_2$ and invoke the strict convexity of F to write

$$\begin{aligned} F\big(\boldsymbol{\nu}(\tilde{f} + \tau q_1 + (1-\tau)q_2), \mathbf{y}\big) &= F\big(\boldsymbol{\nu}(\tau f_1 + (1-\tau) f_2), \mathbf{y}\big) \\ &< \tau F\big(\boldsymbol{\nu}(f_1), \mathbf{y}\big) + (1-\tau) F\big(\boldsymbol{\nu}(f_2), \mathbf{y}\big) \\ &< \tau F\big(\boldsymbol{\nu}(\tilde{f} + q_1), \mathbf{y}\big) + (1-\tau) F\big(\boldsymbol{\nu}(\tilde{f} + q_1), \mathbf{y}\big). \end{aligned}$$

This shows that the data term of J is strictly convex in q when w is fixed, while the same applies if we switch the role of the components. Since the regularization term is strictly convex in w, this implies that J is strictly convex in (w, q) and hence in $f \in \mathcal{H}_{\mathrm{L}}$ by linearity.

(iii) Continuity of J: It simply follows for the fact that all the underlying operators and functionals are continuous.

Properties (i), (ii) and (iii) ensure that the minimizer f_0 of $J(\cdot|\mathbf{y},\lambda)$ over $\mathcal{H}_{\mathrm{L}}$ exists and is unique (by Theorem 36). To obtain its parametric form, we define the constants $\mathbf{y}_0 = \boldsymbol{\nu}(f_0) \in \mathbb{R}^M$ and $\mathbf{c}_0 = (c_1, \ldots, c_{N_0}) = \boldsymbol{\phi}(f_0) \in \mathbb{R}^{N_0}$. Using the property that $\mathcal{H}_{\mathrm{L}}$ is the direct sum of $\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$ and $\mathcal{N}_{\mathrm{L}}$, we then rewrite the solution as $f_0 = \tilde{f}_0 + q_0$ with $q_0 = \mathrm{Proj}_{\mathcal{N}_{\mathrm{L}}}\{f_0\} \in \mathcal{N}_{\mathrm{L}}$ and $\tilde{f}_0 = f_0 - q_0 \in \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$. Since $\boldsymbol{\phi}(\tilde{f}_0) = \mathbf{0}$ as a result of this projection, we have $\boldsymbol{\phi}(q_0) = \mathbf{c}_0$, which implies that

$$q_0 = \sum_{n=1}^{N_0} c_n p_n.$$

Similarly, $\boldsymbol{\nu}(\tilde{f}_0) = \mathbf{y}_0 - \boldsymbol{\nu}(q_0)$, which allows us to specify $\tilde{f}_0$ as the solution of the generalized interpolation problem

$$\arg \min_{\tilde{f} \in \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}} \|\tilde{f}\|_{\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}}^2 \quad \text{s.t.} \quad \boldsymbol{\nu}(\tilde{f}) = \mathbf{y}_0 - \boldsymbol{\nu}(q_0)$$

with $\|\tilde{f}\|_{\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}}^2 = \|\mathrm{L}\tilde{f}\|_{L_2(\mathbb{R}^d)} = \|\mathrm{L}\{\tilde{f}+q_0\}\|_{L_2(\mathbb{R}^d)}$. We then invoke the abstract representer theorem (Theorem 18) which tells us that $\tilde{f}_0 \in \mathrm{span}\{\nu_m^*\}_{m=1}^M$ where ν_m^* is the corresponding Riesz conjugate of ν_m. The combination of these elements yields the parametric expansion

$$f_0 = \sum_{m=1}^{M} a_m \mathrm{A}_{\boldsymbol{\phi}}\{\nu_m\} + \sum_{n=1}^{N_0} c_n p_n$$

where the operator $\mathrm{A}_{\boldsymbol{\phi}}$ (see Theorem 13) is the Riesz map $\mathcal{H}'_{\mathrm{L},\boldsymbol{\phi}} \to \mathcal{H}_{\mathrm{L},\boldsymbol{\phi}}$.

The reverse Riesz map $\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}} \to \mathcal{H}'_{\mathrm{L},\boldsymbol{\phi}}$ is the operator $(\mathrm{L}^*\mathrm{L})$, which can be applied to $\tilde{f}_0$ to obtain the conjugate function

$$\tilde{f}_0^* = (\mathrm{L}^*\mathrm{L})\{\tilde{f}_0\} = \sum_{m=1}^{M} a_m \nu_m,$$

which is included in $\mathcal{H}'_{\mathrm{L},\boldsymbol{\phi}}$ by definition (see functional mapping in Figure 2). Hence, we necessarily have that $\boldsymbol{p}(\tilde{f}_0^*) = \mathbf{0}$, which translates into the stated orthogonality property. Finally, we rewrite the solution as

$$\begin{aligned} f_0 &= \mathrm{A}_{\boldsymbol{\phi}}\{\tilde{f}_0^*\} + \sum_{n=1}^{N_0} c_n p_n \\ &= \mathrm{A}\{\tilde{f}_0^*\} + \sum_{n=1}^{N_0} \big(\langle \mathrm{A}\{\phi_n\}, \tilde{f}_0^* \rangle + c_n\big) p_n \\ &= \sum_{m=1}^{M} a_m \mathrm{A}\{\nu_m\} + \sum_{n=1}^{N_0} b_n p_n \end{aligned}$$

where we have made use of Property 5 in Theorem 13 to readjust the constants associated with the null-space component.

□

3.3 Discretization and numerical solutions

Besides the guarantee of unicity, the remarkable outcome of Theorem 20 is that the generic form of the solution (112) is a linear combination of the basis vectors $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_M)$ and $\boldsymbol{p} = (p_1, \ldots, p_{N_0})$ where the φ_m are specified by (113). Consequently, we can obtain an exact discretization of the problem by searching for the optimal solution within the finite-dimensional reconstruction space

$$\mathcal{V}_{\mathrm{L},\boldsymbol{\nu}} = \{g = \boldsymbol{\varphi}^T \mathbf{a} + \boldsymbol{p}^T \mathbf{b} : \mathbf{a} \in \mathbb{R}^M \text{ and } \mathbf{b} \in \mathbb{R}^{N_0}\}. \tag{114}$$

In other words, we have that

$$\arg\min_{f \in \mathcal{H}_{\mathrm{L}}} J(f|\mathbf{y}, \lambda) = \arg\min_{g \in \mathcal{V}_{\mathrm{L},\boldsymbol{\nu}}} J(g|\mathbf{y}, \lambda)$$

where the minimization on the left-hand side converts the original continuous-domain problem into a finite-dimensional optimization in terms of the parameter vectors $\mathbf{a}$ and $\mathbf{b}$.

To set up the numerical problem, we need to express $J(g|\mathbf{y},\lambda) = F(\boldsymbol{\nu}(g),\mathbf{y}) + \lambda\|\mathrm{L}g\|_{L_2}^2$ for $g \in \mathcal{V}_{\mathrm{L},\boldsymbol{\nu}}$ in terms of $(\mathbf{a},\mathbf{b})$. Due to the specific form (113) of the basis functions and the null-space property $\mathrm{L}\{p_n\} = 0$, we readily find that

$$\begin{aligned}\|\mathrm{L}g\|_{L_2}^2 &= \|\mathrm{L}\{\boldsymbol{\varphi}^T\mathbf{a}\}\|_{L_2}^2 = \langle(\mathrm{L}^*\mathrm{L})\{\boldsymbol{\varphi}^T\mathbf{a}\},\boldsymbol{\varphi}^T\mathbf{a}\rangle \\ &= \langle\boldsymbol{\nu}^T\mathbf{a},\boldsymbol{\varphi}^T\mathbf{a}\rangle = \mathbf{a}^T\mathbf{G}\mathbf{a}\end{aligned}$$

where $\mathbf{G} = \langle\boldsymbol{\nu},\boldsymbol{\varphi}^T\rangle$ is a symmetric matrix of size M whose the entries are given by

$$[\mathbf{G}]_{m,m'} = \langle\nu_m,\varphi_{m'}\rangle = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\nu_m(\boldsymbol{x})G_{\mathrm{L}^*\mathrm{L}}(\boldsymbol{x},\boldsymbol{y})\nu_{m'}(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \tag{115}$$

The latter can also be written as $[\mathbf{G}]_{m,m'} = \langle\nu_m,\mathrm{A}\{\nu_{m'}\}\rangle$ where A is a positive operator (?), which implies that the matrix $\mathbf{G}$ is positive-definite (see Appendix A). Likewise, by invoking the linearity of the measurement operator $\boldsymbol{\nu}:\mathcal{H}_{\mathrm{L}}\to\mathbb{R}^M$, we find that

$$\boldsymbol{\nu}(g) = \mathbf{G}\mathbf{a} + \mathbf{P}\mathbf{b}$$

where $\mathbf{G}$ is same as before and where $\mathbf{P}$ is a matrix of size $M\times N_0$ whose entries are given by

$$[\mathbf{P}]_{m,n} = \langle\nu_m,p_n\rangle = \int_{\mathbb{R}^d}\nu_m(\boldsymbol{x})p_n(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{116}$$

To sum up, given the error function $F:\mathbb{R}^M\times\mathbb{R}^M\to\mathbb{R}$ and the data point $\mathbf{y}\in\mathbb{R}^M$, we restate the optimization problem (111) as

$$\arg\min_{\mathbf{a}\in\mathbb{R}^M,\,\mathbf{b}\in\mathbb{R}^{N_0}}\left\{F(\mathbf{G}\mathbf{a}+\mathbf{P}\mathbf{b},\mathbf{y}) + \lambda\,\mathbf{a}^T\mathbf{G}\mathbf{a}\right\} \tag{117}$$

where the corresponding "sensing" matrices $\mathbf{G}\in\mathbb{R}^{M\times M}$ and $\mathbf{P}\in\mathbb{R}^{M\times N_0}$ are defined as

$$\begin{aligned}\mathbf{G} &= \left[\boldsymbol{\nu}(\varphi_1)\ \boldsymbol{\nu}(\varphi_2)\ \cdots\ \boldsymbol{\nu}(\varphi_M)\right] \\ \mathbf{P} &= \left[\boldsymbol{\nu}(p_1)\ \cdots\ \boldsymbol{\nu}(p_{N_0})\right]\end{aligned}$$

and where $\lambda\in\mathbb{R}^+$ is our adjustable regularization parameter. Since F is convex, we can then solve (117) iteratively by applying a steepest-descent algorithm or a variant thereof. This requires the specification of the gradient of the loss functional

$$\frac{\partial F(\mathbf{z},\mathbf{y})}{\partial\mathbf{z}} = \boldsymbol{\nabla}F(\mathbf{z},\mathbf{y})$$

and the choice of an appropriate step size $\tau \in \mathbb{R}^+$. Starting from an arbitrary initialization $(\mathbf{a}_0, \mathbf{b}_0)$, the values of $(\mathbf{a}, \mathbf{b})$ are then updated recursively according to the formulas

$$\mathbf{a}_{k+1} = \mathbf{a}_k - \tau \mathbf{G} \Big(\boldsymbol{\nabla} F(\mathbf{G}\mathbf{a}_k + \mathbf{P}\mathbf{b}_k, \mathbf{y}) + 2\lambda \mathbf{a}_k \Big)$$
$$\mathbf{b}_{k+1} = \mathbf{b}_k - \tau \mathbf{P}^T \boldsymbol{\nabla} F(\mathbf{G}\mathbf{a}_k + \mathbf{P}\mathbf{b}_k, \mathbf{y})$$

until a suitable stopping criterion is met. This algorithm is guaranteed to converge to some fixed point $(\mathbf{a}_*, \mathbf{b}_*)$ as $k \to \infty$ provided that $\tau > 0$ be taken sufficiently small. The exact (and unique) continuous-domain solution of the problem is then obtained by injecting these coefficients in the expansion formula (112).

3.3.1 Least-squares approximation problems

For demonstration purposes, we consider the generalized smoothing spline problem where the loss functional is the least-squares criterion $\|\boldsymbol{\nu}(g) - \mathbf{y}\|_2^2$. The optimal solution is then specified as

$$J_{\mathrm{LS}}(\mathbf{a}, \mathbf{b} | \mathbf{y}, \lambda) = \|\mathbf{G}\mathbf{a} + \mathbf{P}\mathbf{b} - \mathbf{y}\|_2^2 + \lambda \, \mathbf{a}^T \mathbf{G} \mathbf{a} \tag{118}$$
$$\arg \min_{(\mathbf{a},\mathbf{b}) \in \mathbb{R}^{M+N_0}} J_{\mathrm{LS}}(\mathbf{a}, \mathbf{b} | \mathbf{y}, \lambda) \tag{119}$$

By using standard differential calculus and the property that $\mathbf{G} = \mathbf{G}^T$, we first partially differentiate $J_{\mathrm{LS}}(\mathbf{a}, \mathbf{b} | \mathbf{y}, \lambda)$ with respect to $\mathbf{a}$, which gives

$$\begin{aligned} \frac{\partial J_{\mathrm{LS}}(\mathbf{a}, \mathbf{b} | \mathbf{y}, \lambda)}{\partial \mathbf{a}} &= 2\mathbf{G}(\mathbf{G}\mathbf{a} + \mathbf{P}\mathbf{b} - \mathbf{y}) + 2\lambda \mathbf{G}\mathbf{a}, \\ &= 2\mathbf{G}\Big((\mathbf{G} + \lambda \mathbf{I})\mathbf{a} + \mathbf{P}\mathbf{b} - \mathbf{y} \Big). \end{aligned}$$

This leads to the identification of the first condition of optimality

$$(\mathbf{G} + \lambda \mathbf{I})\mathbf{a} + \mathbf{P}\mathbf{b} = \mathbf{y} \tag{120}$$

which forces the above partial derivative to vanish. Similarly, we calculate the partial derivative of $J_{\mathrm{LS}}(\mathbf{a}, \mathbf{b} | \mathbf{y}, \lambda)$ with respect to $\mathbf{b}$ and set it to zero as

$$\frac{\partial J_{\mathrm{LS}}(\mathbf{a}, \mathbf{b} | \mathbf{y}, \lambda)}{\partial \mathbf{b}} = 2\mathbf{P}^T(\mathbf{G}\mathbf{a} + \mathbf{P}\mathbf{b} - \mathbf{y}) = 0$$

Upon substitution of the value of $\mathbf{y}$ given by (120), this provides us with a second equation

$$\mathbf{P}^T \mathbf{a} = \mathbf{0} \tag{121}$$

which is equivalent to the "orthogonality" property in Theorem 20. The optimal spline is then found by jointly solving (120) and (121), which results in the closed-form solution

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} (\mathbf{G}+\lambda\mathbf{I}) & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}.$$

For the limit case where $\lambda \to 0$, we get the generalized interpolant $f_{(0)} \in \mathcal{H}_{\mathrm{L}}$ with expansion coefficients $(\mathbf{a}_{(0)}, \mathbf{b}_{(0)})$, which is such that $\boldsymbol{\nu}(f_{(0)}) = \mathbf{y}$. Based on (120) and (121), it is then possible to simplify the corresponding spline energy as

$$\|\mathrm{L}f_{(0)}\|_{L_2}^2 = \mathbf{a}_{(0)}^T\mathbf{G}\mathbf{a}_{(0)} = \mathbf{a}_{(0)}^T(\mathbf{y}-\mathbf{P}\mathbf{b}_{(0)}) = \mathbf{a}_{(0)}^T\mathbf{y} - \mathbf{b}_{(0)}^T(\mathbf{P}^T\mathbf{a}_{(0)}) = \mathbf{a}_{(0)}^T\mathbf{y}.$$

At the other extreme for $\lambda \to \infty$, we have that

$$\begin{pmatrix} \mathbf{a}_{(\infty)} \\ \mathbf{b}_{(\infty)} \end{pmatrix} = \lim_{\lambda\to\infty} \begin{pmatrix} \lambda\mathbf{I} & \mathbf{P} \\ \mathbf{P}^T & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}$$

This yields $\mathbf{a}_{(\infty)} = \mathbf{0}$ and

$$\mathbf{b}_{(\infty)} = (\mathbf{P}^T\mathbf{P})^{-1}\mathbf{P}^T\mathbf{y}$$

where the latter represents the orthogonal projection of $\mathbf{y}$ in the space spanned by the column vectors of $\mathbf{P}$.

3.3.2 Specific examples

3.3.3 Generalized interpolation revisited

The orthogonality property $\mathbf{P}^T\mathbf{a} = \mathbf{0}$ introduces a linear dependency between the expansion coefficients $\mathbf{a} \in \mathbb{R}^M$ and $\mathbf{b} \in \mathbb{R}^{N_0}$ in (114). This suggests that the search space actually only has M degrees of freedom. Since the optimal solution $f_{(\lambda)}$ of the generic optimization problem in Theorem 20 is uniquely determined by its measurement values $\boldsymbol{\nu}(f_{(\lambda)})$, it suffices to identify the underlying basis functions for the generalized interpolation problem

$$\min_{f\in\mathcal{H}_{\mathrm{L}}} \|\mathrm{L}f\|_{L_2(\mathbb{R}^d)}^2 \text{ s.t. } \boldsymbol{\nu}(f) = \mathbf{y}.$$

We shall do so by using the fact that $\mathcal{H}_{\mathrm{L}}$ is a RKHS with respect to the inner product given in Theorem 12. Without loss of generality, we assume that the first N_0 measurement functionals are linearly independent with respect to $\mathcal{N}_{\mathrm{L}}$ to take advantage of the last property in Proposition 16.

Let $\mathbf{y}_0 = (y_1, \ldots, y_{N_0})$ and $\boldsymbol{\nu}_0 = (\nu_1, \ldots, \nu_{N_0})$. Since the cross-product matrix $\mathbf{P}_0 \in \mathbb{R}^{N_0 \times N_0}$ with

$$\mathbf{P}_0 = [\boldsymbol{\nu}_0(p_1) \ \cdots \ \boldsymbol{\nu}_0(p_{N_0})]$$

is invertible by hypothesis, we construct $\boldsymbol{\phi}_0 = (\phi_1, \ldots, \phi_{N_0}) = \mathbf{P}_0^{-1}\boldsymbol{\nu}_0$, which yields the unique biorthogonal system $\{\phi_n, p_n\}_{n=1}^{N_0}$ with the property $\phi_n \in \text{span}\{\nu_m\}_{n=1}^{N_0}$. Within that framework, any $f \in \mathcal{H}_{\text{L}} = \mathcal{H}_{\text{L},\boldsymbol{\phi}_0} \oplus \mathcal{N}_{\text{L}}$ has a unique decomposition as $f = q + g$ with $q = \text{Proj}_{\mathcal{N}_{\text{L}}}\{f\} \in \mathcal{N}_{\text{L}}$ and $g = f - q \in \mathcal{H}_{\text{L},\boldsymbol{\phi}_0}$. Moreover, $\boldsymbol{\nu}_0(g) = \mathbf{P}_0\boldsymbol{\phi}_0(g) = \mathbf{0}$ by construction, so that we can transfer the constraint $\boldsymbol{\nu}_0(f) = \mathbf{y}_0$ to the null-space component. This results in the basic interpolation problem

$$q_0 = q \in \mathcal{N}_{\text{L}} : \boldsymbol{\nu}_0(q) = \mathbf{y}_0$$

the solution of which is simply

$$q_0 = \sum_{n=1}^{N_0} c_n p_n$$

with $\mathbf{c}_0 = (c_1, \cdots, c_{N_0}) = \mathbf{P}_0^{-1}\mathbf{y}_0$.

Next, we recall that $\mathcal{H}_{\text{L},\boldsymbol{\phi}_0}$ is a Hilbert space equipped with the inner-product $\langle g_1, g_2 \rangle_{\text{L}} = \langle \text{L}g_1, \text{L}g_2 \rangle$ (see Theorem 10). Defining $\mathbf{y}_1 = (y_{N_0+1}, \ldots, y_M)$ and $\boldsymbol{\nu}_1 = (\nu_{N_0+1}, \ldots, \nu_M)$ so that $\mathbf{y} = (\mathbf{y}_0, \mathbf{y}_1)$ and $\boldsymbol{\nu} = (\boldsymbol{\nu}_0, \boldsymbol{\nu}_1)$, we thereby reformulate our optimization problem in the decoupled form

$$g_1 = \min_{g \in \mathcal{H}_{\text{L},\boldsymbol{\phi}_0}} \|g\|_{\text{L}}^2 \text{ s.t. } \boldsymbol{\nu}_1(g) = \mathbf{y}_1 - \boldsymbol{\nu}_1(q_0).$$

which falls within the Hilbert-space framework of Theorem 18. The minimum-norm solution then takes the standard parametric form

$$g_1 = \sum_{m=N_0+1}^{M} c_m \nu_m^*$$

with $\nu_m^* = \text{A}_{\boldsymbol{\phi}_0}\{\nu_m\}$ where $\text{A}_{\boldsymbol{\phi}_0}$ is the Riesz map $\mathcal{H}'_{\text{L},\boldsymbol{\phi}_0} \to \mathcal{H}_{\text{L},\boldsymbol{\phi}_0}$ whose (reproducing) kernel is specified in Theorem 11. By enforcing the interpolation constraint, we find that the expansion coefficients of g_1 are given by

$$\mathbf{c}_1 = (c_{N_0+1}, \cdots, c_M) = \mathbf{G}_1^{-1}\big(\mathbf{y}_1 - \mathbf{P}_1\mathbf{c}_0\big)$$

where the underlying sensing matrices are

$$\mathbf{P}_1 = [\boldsymbol{\nu}_1(p_1) \ \cdots \ \boldsymbol{\nu}_1(p_{N_0})]$$
$$\mathbf{G}_1 = [\boldsymbol{\nu}_1(\nu^*_{N_0+1}) \ \cdots \ \boldsymbol{\nu}_1(\nu^*_M)].$$

The solution of the original problem is the generalized spline interpolant $f = q_0 + g_1$, which can be written as

$$f = \sum_{n=1}^{N_0} c_n p_n + \sum_{m=N_0+1}^{M} c_m \mathrm{A}_{\boldsymbol{\phi}_0}\{\nu_m\}. \tag{122}$$

The bottom line is that f lies in a subspace of dimension M and that it can be parametrized in terms of its measurements $\mathbf{y} = \boldsymbol{\nu}(f) \in \mathbb{R}^M$. Specifically, the linear, one-to-one relation between $\mathbf{y} = (\mathbf{y}_0, \mathbf{y}_1)$ and the expansion coefficients of f in (122) is summarized by the matrix equation

$$\mathbf{c} = \begin{pmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{P}_0^{-1} & \mathbf{0} \\ -\mathbf{G}_1^{-1}\mathbf{P}_1\mathbf{P}_0^{-1} & \mathbf{G}_1^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{pmatrix},$$

while the corresponding spline energy is

$$\|\mathrm{L}f\|_{L_2}^2 = \|\mathrm{L}g_1\|_{L_2}^2 = \mathbf{c}_1^T \mathbf{G}_1 \mathbf{c}_1 = \mathbf{c}_1^T(\mathbf{y}_1 - \mathbf{P}_1\mathbf{c}_0).$$

3.4 Epilogue: back to the finite-dimensional world

We have used the term "representer theorem" to convey the remarkable property that the continuous-domain minimizer of any convex cost functional with a quadratic regularization term lives in a fixed finite-dimensional space that solely depends on the type of measurements (e.g. the operator $\boldsymbol{\nu}$) and the regularization operator L. We have also discussed the practical benefit of this reduction of dimensionality as it results in an exact discretization where the problem is recast as a finite-dimensional numerical optimization program.

To close the topic of functional minimization, we shall now adopt a purely discrete point of view and provide the finite-dimensional counterpart of Theorem 20; that is, the representer theorem for linear inverse problems with Tikhonov (or ℓ_2) regularization. The task there is to recover an unknown vector $\mathbf{c} \in \mathbb{R}^N$ (our discrete signal) from a noisy set of $M < N$ linear measurements $y_m = \langle \mathbf{c}, \mathbf{h}_m \rangle + \epsilon_m$ where ϵ_m is some unknown/random disturbance component.

The recovery is done by minimizing the discrepancy (fitting error) between the true measurements $\mathbf{y}$ and the predicted ones—as quantified by

$F(\mathbf{Hc}, \mathbf{y})$—subject to a regularization constraint on $\mathbf{c}$. Specifically, the entities that enter the formulation are:

- the finite-dimensional signal $\mathbf{c} \in \mathbb{R}^N$ to be recovered by the algorithm;
- the input data vector $\mathbf{y} \in \mathbb{R}^M$ with $M < N$;
- the system matrix $\mathbf{H} = [\mathbf{h}_1 \dots \mathbf{h}_M]^T \in \mathbb{R}^{M \times N}$;
- the error functional $F : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}^+$, which is assumed to be convex and coercive;
- the regularization operator specified by the matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$ of rank $N' \leq N$;
- the null space of $\mathbf{L}$: $\mathcal{N}_{\mathbf{L}} = \text{span}\{\mathbf{p}_n\}_{n=1}^{N_0}$ with $0 \leq N_0 = N - N'$;
- the RKHS $\mathcal{H} \subseteq \mathbb{R}^N$ associated the inner product $\langle \mathbf{c}_1, \mathbf{c}_2 \rangle_{\mathcal{H}} = \langle \mathbf{L}\mathbf{c}_1, \mathbf{L}\mathbf{c}_2 \rangle$ and the reproducing kernel $\mathbf{R} = (\mathbf{L}^T\mathbf{L})^\dagger$ (see Section 2.3 and Proposition 5);
- The adjustable regularization parameter $\lambda \in \mathbb{R}^+$.

Similar to the continuous-domain setting, we assume that the recovery problem is well-posed over the null space of $\mathbf{L}$. This is equivalent to the existence of a constant $c > 0$ such that $\|\mathbf{p}\|_2 \leq c\|\mathbf{Hp}\|_2$ for all $\mathbf{p} \in \mathcal{N}_{\mathbf{L}}$ so that $\mathbf{Hp} = \mathbf{0} \Leftrightarrow \mathbf{p} = \mathbf{0}$.

Theorem 21 (Discrete representer theorem)**.** *Let $\{\mathbf{p}_m\}_{n=1}^{N_0}$ with $N_0 < N$ be an orthonormal basis of $\mathcal{N}_{\mathbf{L}}$ with corresponding system matrix $\mathbf{P} \in \mathbb{R}^{M \times N_0}$ of rank N_0 (condition for a proper regularization) with $[\mathbf{P}]_{m,n} = \langle \mathbf{h}_m, \mathbf{p}_n \rangle$. Then, the generic convex minimization problem*

$$\mathbf{c}_{\text{opt}} = \arg \min_{\mathbf{c} \in \mathbb{R}^M} \left(F(\mathbf{Hc}, \mathbf{y}) + \lambda \|\mathbf{Lc}\|_2^2 \right) \tag{123}$$

has a unique solution solution $\mathbf{c}_{\text{opt}}$, which lies in a finite-dimensional subspace of dimension M fully determined by $\mathbf{H}$ and $\mathbf{L}$. Specifically, there is a unique set of coefficients $\mathbf{a} = (a_m) \in \mathbb{R}^M$ and $\mathbf{b} = (b_n) \in \mathbb{R}^{N_0}$ such that

$$\mathbf{c}_{\text{opt}} = \sum_{m=1}^{M} a_m \widetilde{\mathbf{h}}_m + \sum_{n=1}^{N_0} b_n \mathbf{p}_n$$

with $\widetilde{\mathbf{h}}_m = \mathbf{R}\mathbf{h}_m$, subject to the orthogonality constraint $\mathbf{P}^T\mathbf{a} = \mathbf{0}$ which restricts the effective number of degrees of freedom to M.

The key element for the proof is the orthogonal projection operator $\text{Proj}_{\mathcal{N}_\mathbf{L}} : \mathbb{R}^N \to \mathcal{N}_\mathbf{L}$, which is specified by $\mathbf{c} \mapsto \sum_{n=1}^{N_0} \langle \mathbf{p}_n, \mathbf{c} \rangle \mathbf{p}_n$. Consequently, any element $\mathbf{c} \in \mathbb{R}^N$ has a unique expansion as $\mathbf{c} = \mathbf{p}^\perp + \mathbf{p}$ with $\mathbf{p} = \text{Proj}_{\mathcal{N}_\mathbf{L}}\{\mathbf{c}\} \in \mathcal{N}_\mathbf{L}$ and $\mathbf{p}^\perp = \mathbf{c} - \mathbf{p} \in \mathcal{H}$ with the property that $\langle \mathbf{p}^\perp, \mathbf{p} \rangle = 0$. In other words, we have the direct sum decomposition $\mathbb{R}^N = \mathcal{H} \oplus \mathcal{N}_\mathbf{H}$ which allows us to replicate the proof of Theorem 20 with the subsequent list of substitutions:

- Regularization functional : $\langle \mathrm{L}f, \mathrm{L}f \rangle \quad \rightsquigarrow \quad \langle \mathbf{Lc}, \mathbf{Lc} \rangle$
- Measurement operator: $f \mapsto \boldsymbol{\nu}(f) \quad \rightsquigarrow \quad \mathbf{c} \mapsto \mathbf{Hc}$
- RKHS associated with the regularization: $\mathcal{H}_{\mathrm{L},\boldsymbol{\phi}} \quad \rightsquigarrow \quad \mathcal{H}$
- Riesz map: $\mathrm{A}_{\boldsymbol{\phi}} = \mathrm{L}_{\boldsymbol{\phi}}^{-1} \mathrm{L}_{\boldsymbol{\phi}}^{-1*} \quad \rightsquigarrow \quad \mathbf{R} = \mathbf{R}^{1/2}\mathbf{R}^{1/2}$
- Stable right-inverse: $\mathrm{L}_{\boldsymbol{\phi}}^{-1} \quad \rightsquigarrow \quad \mathbf{R}^{1/2}$.

Once more, the solution is composed of two terms: a primary part, whose parametric form follows from the abstract representation theorem (Theorem 18), and a secondary null-space component that does not affect the regularization cost. The orthogonality condition ensures that the latter component contributes maximally to the reduction of the fitting error (data term).

The simplest instance of (123) is the regularized least-squares problem

$$\arg \min_{\mathbf{c} \in \mathbb{R}^M} \left(\|\mathbf{Hc} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{Lc}\|_2^2 \right),$$

which admits the well-known closed-form solution

$$\mathbf{c}_{\text{opt}} = (\mathbf{H}^T\mathbf{H} + \lambda \mathbf{L}^T\mathbf{L})^{-1}\mathbf{H}^T\mathbf{y}. \tag{124}$$

To show that the compatibility of this classical formula with the expansion in Theorem 21, let us assume, for simplicity, that the regularization operator is invertible (i.e., $N_0 = 0$ and $\mathbf{R} = (\mathbf{L}^T\mathbf{L})^{-1}$). This allows us to write the following sequence of (equivalent) identities

$$\begin{aligned}
(\mathbf{H}^T\mathbf{H})\mathbf{R}\mathbf{H}^T + \lambda \mathbf{H}^T &= (\mathbf{H}^T\mathbf{H})\mathbf{R}\mathbf{H}^T + \lambda (\underbrace{\mathbf{L}^T\mathbf{L}\mathbf{R}}_{=\mathbf{I}_N})\mathbf{H}^T \\
\mathbf{H}^T \left(\mathbf{H}\mathbf{R}\mathbf{H}^T + \lambda \mathbf{I}_M \right) &= \left(\mathbf{H}^T\mathbf{H} + \lambda \mathbf{L}^T\mathbf{L} \right) \mathbf{R}\mathbf{H}^T \\
\left(\mathbf{H}^T\mathbf{H} + \lambda \mathbf{L}^T\mathbf{L} \right)^{-1} \mathbf{H}^T &= \mathbf{R}\mathbf{H}^T \left(\mathbf{H}\mathbf{R}\mathbf{H}^T + \lambda \mathbf{I}_M \right)^{-1},
\end{aligned}$$

By plugging the last equation in (124), we get $\mathbf{c}_{\text{opt}} = \sum_{m=1}^M a_m \widetilde{\mathbf{h}}_m$ with $\mathbf{a} = \left(\mathbf{H}\mathbf{R}\mathbf{H}^T + \lambda \mathbf{I}_M \right)^{-1} \mathbf{y}$, which is the desired form. In fact, if we set $\mathbf{G} =$

$\mathbf{HRH}^T$—the matrix equivalent of (100)—we end up with the same formula as (98), which highlights the parallel between the discrete and continuous forms of the problem.