Kernel Methods and Operator Learning

These notes provide an introduction to kernel learning through four steps:

1. Minimum-norm interpolation in Hilbert spaces

2. Reproducing kernels and the representer theorem

3. Kernel ridge regression (KRR) as a natural generalization

4. The transition from function learning to operator learning

The end briefly mentions how the paper by Batlle, and others, clarifies the role of reproducing kernel Hilbert spaces when applied to operator learning tasks.

1. Minimum-Norm Interpolation in a Hilbert Space

Let $(\mathcal{H}, \langle\cdot,\,\cdot\rangle_{\mathcal{H}})$ be a Hilbert space of real-valued functions on a set Ω. Assume point-evaluation is continuous (true for all classical RKHS/Sobolev spaces considered here).

Given data

$$(x_1, y_1), \ldots, (x_n, y_n), \qquad x_i \in \Omega, \;\; y_i \in \mathbb{R},$$

consider the variational problem

$$\hat{f} = \arg\min_{g \in \mathcal{H}} \big\{ \, \|g\|_{\mathcal{H}} : g(x_i) = y_i, \; i = 1, \ldots, n \big\}. \tag{1}$$

This is *the* classical spline problem: among all interpolants, choose the one of minimal "energy" measured by the Hilbert norm.

A concrete model space: the Sobolev space $H_0^1(0,1)$

For the remainder of this section we take

$$\mathcal{H} = H_0^1(0,1), \qquad \langle f,\, g\rangle_{\mathcal{H}} = \int_0^1 f'(x)\, g'(x)\, dx,$$

the Sobolev space of absolutely continuous functions on $(0,1)$ with square–integrable derivatives and homogeneous boundary conditions $f(0) = f(1) = 0$. This is a Hilbert space, and point evaluation $f \mapsto f(s)$ is a continuous linear functional on $H_0^1(0,1)$ (by the Sobolev embedding $H_0^1(0,1) \hookrightarrow C^0[0,1]$). The norm

$$\|f\|_{\mathcal{H}}^2 = \int_0^1 |f'(x)|^2\, dx$$

encodes the "bending energy" of the interpolant, making H_0^1 the natural setting for classical spline theory.

To solve the optimization problem, introduce Lagrange multipliers $\{\lambda_i\}_{i=1}^n$ and define the augmented functional

$$\mathcal{L}(f, \lambda_1, \ldots, \lambda_n) = \tfrac{1}{2} \int_0^1 |f'(x)|^2\, dx \; + \; \sum_{i=1}^n \lambda_i \big(f(x_i) - y_i\big).$$

Let $\varphi \in H_0^1(0,1)$ be an arbitrary variation. The first variation of $\mathcal{L}$ in the direction φ is

$$\frac{d}{d\varepsilon}\Big|_{\varepsilon=0} \mathcal{L}(f + \varepsilon\varphi, \lambda_1, \ldots, \lambda_n) = \int_0^1 f'(x)\,\varphi'(x)\,dx \;+\; \sum_{i=1}^{n} \lambda_i\,\varphi(x_i).$$

Integrating by parts and using $\varphi(0) = \varphi(1) = 0$ gives

$$\int_0^1 f'(x)\,\varphi'(x)\,dx = -\int_0^1 f''(x)\,\varphi(x)\,dx,$$

so the stationarity condition $\delta\mathcal{L}(f,\lambda)[\varphi] = 0$ for all φ becomes

$$-\int_0^1 f''(x)\,\varphi(x)\,dx \;+\; \sum_{i=1}^{n} \lambda_i\,\varphi(x_i) = 0, \qquad \forall \varphi \in H_0^1(0,1).$$

Recall that point evaluation can be written in terms of Dirac deltas:

$$\varphi(x_i) = \int_0^1 \delta(x - x_i)\,\varphi(x)\,dx.$$

Hence

$$0 = \int_0^1 \Big(-f''(x) + \sum_{i=1}^{n} \lambda_i\,\delta(x - x_i) \Big)\,\varphi(x)\,dx \qquad \forall \varphi.$$

Since this holds for all test functions φ, we obtain the Euler–Lagrange equation in the sense of distributions:

$$-\frac{d^2}{dx^2}\hat{f}(x) = \sum_{i=1}^{n} \lambda_i\,\delta(x - x_i), \qquad \hat{f}(0) = \hat{f}(1) = 0.$$

Renaming $\alpha_i := \lambda_i$ we recover

$$-\frac{d^2}{dx^2}\hat{f}(x) = \sum_{i=1}^{n} \alpha_i\,\delta(x - x_i).$$

Green's function of $-d^2/dx^2$ on $(0,1)$. The Green's function $G(x,s)$ for the operator $-d^2/dx^2$ with homogeneous Dirichlet boundary conditions is defined as the unique function satisfying

$$-\frac{\partial^2}{\partial x^2} G(x,s) = \delta(x - s), \qquad G(0,s) = G(1,s) = 0.$$

For fixed $s \in (0,1)$, away from $x = s$ we have

$$-\frac{\partial^2}{\partial x^2} G(x,s) = 0,$$

so $G(\cdot, s)$ is linear on each interval $(0,s)$ and $(s,1)$. Write

$$G(x,s) = \begin{cases} a_1(s)\,x + b_1(s), & 0 \le x \le s, \\ a_2(s)\,x + b_2(s), & s \le x \le 1. \end{cases}$$

The boundary conditions give

$$G(0,s) = 0 \Rightarrow b_1(s) = 0, \qquad G(1,s) = 0 \Rightarrow a_2(s) + b_2(s) = 0.$$

We also impose continuity at $x = s$:

$$G(s^-, s) = G(s^+, s) \quad \Rightarrow \quad a_1(s)\, s = a_2(s)\, s + b_2(s).$$

Finally, integrating the defining equation across a small interval $(s-\varepsilon, s+\varepsilon)$ and letting $\varepsilon \to 0$ enforces the jump condition in the derivative:

$$\int_{s-\varepsilon}^{s+\varepsilon} -\frac{\partial^2}{\partial x^2} G(x, s)\, dx = \int_{s-\varepsilon}^{s+\varepsilon} \delta(x - s)\, dx = 1,$$

so

$$-\Big(\partial_x G(s^+, s) - \partial_x G(s^-, s)\Big) = 1 \quad \Rightarrow \quad a_1(s) - a_2(s) = 1.$$

Solving this linear system:

$$\begin{aligned} b_1(s) &= 0, \\ b_2(s) &= -a_2(s), \\ a_1(s)\, s &= a_2(s)\, s + b_2(s) = a_2(s)(s-1), \\ a_1(s) - a_2(s) &= 1. \end{aligned}$$

From the third relation,

$$a_1(s) = a_2(s)\, \frac{s-1}{s},$$

and substituting into $a_1(s) - a_2(s) = 1$ gives

$$a_2(s)\Big(\frac{s-1}{s} - 1\Big) = 1 \quad \Rightarrow \quad a_2(s)\frac{-1}{s} = 1 \quad \Rightarrow \quad a_2(s) = -s,$$

so

$$a_1(s) = 1 - s, \qquad b_2(s) = -a_2(s) = s.$$

Thus

$$G(x, s) = \begin{cases} (1-s)\, x, & 0 \le x \le s, \\ s(1-x), & s \le x \le 1. \end{cases}$$

Equivalently,

$$G(x, s) = \begin{cases} x(1-s) & x \le s, \\ s(1-x) & x \ge s, \end{cases}$$

which is symmetric in (x, s): $G(x, s) = G(s, x)$.

Representation of the minimizer and emergence of the kernel. Since $-\hat{f}'' = \sum_{i=1}^n \alpha_i \delta(x - x_i)$ with homogeneous Dirichlet boundary conditions, standard Green's function theory gives

$$\hat{f}(x) = \sum_{i=1}^n \alpha_i\, G(x, x_i).$$

The interpolation constraints $\hat{f}(x_j) = y_j$ for $j = 1, \ldots, n$ determine the coefficients $\{\alpha_i\}$ through the linear system

$$\sum_{i=1}^n \alpha_i\, G(x_j, x_i) = y_j, \qquad j = 1, \ldots, n.$$

In this way the Green's function

$$K(x, s) := G(x, s)$$

acts as the reproducing kernel for the Hilbert space $\mathcal{H}$ endowed with the norm $\|f\|_{\mathcal{H}}^2 = \int_0^1 |f'(x)|^2\, dx$. And in the process, we rediscover how to perform linear interpolation.

2. Reproducing Kernels and the Representer Theorem

A Hilbert space $\mathcal{H}$ of functions on a domain Ω is called a *reproducing kernel Hilbert space* (RKHS) if there exists a function

$$K : \Omega \times \Omega \to \mathbb{R}$$

such that:

1. For every $x \in \Omega$, the section $K(\cdot, x) \in \mathcal{H}$.
2. (**Reproducing property**) For all $f \in \mathcal{H}$ and all $x \in \Omega$,

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

The existence of such a kernel determines *all* geometry of $\mathcal{H}$: inner products, norms, orthogonality, and the behavior of point-evaluation functionals.

Representer Theorem

Let $L : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be an empirical loss. That is, L is lower semicontinuous as a function of its n arguments, depends only on evaluations through

$$L(f) = L\big(f(x_1), \ldots, f(x_n)\big) \quad \text{for all } f \in \mathcal{H}.$$

and satisfies pointwise invariance: if $f, g \in \mathcal{H}$ satisfy $f(x_i) = g(x_i)$ for all i, then

$$L(f) = L(g).$$

Furthermore, let

$$\Phi : [0, \infty) \to \mathbb{R} \cup \{+\infty\}$$

be any strictly increasing and lower semicontinuous function of the Hilbert norm $\|f\|_{\mathcal{H}}$.

Now, consider the regularized problem

$$\min_{f \in \mathcal{H}} \Big\{ L\big(f(x_1), \ldots, f(x_n)\big) + \Phi(\|f\|_{\mathcal{H}}) \Big\}.$$

The Representer Theorem states that every minimizer must lie in the n–dimensional subspace

$$\operatorname{span}\big\{K(\cdot, x_1),\, \ldots,\, K(\cdot, x_n)\big\}.$$

Thus, the optimal solution has the finite expansion

$$f(x) = \sum_{i=1}^{n} \alpha_i\, K(x, x_i), \qquad \alpha_i \in \mathbb{R}.$$

This conclusion is *independent* of the choice of loss L or penalty Φ—the structure comes entirely from the geometry of $\mathcal{H}$. (See B. Schölkopf and A. Smola, Learning with Kernels, MIT Press, 2002, Theorem 4.2 for more details about the Representer Theorem)

How the representer theorem relates to the original model problem. In the model space

$$\mathcal{H} = H_0^1(0,1), \qquad \langle f,\, g\rangle_{\mathcal{H}} = \int_0^1 f'(x)\, g'(x)\, dx,$$

two standard facts connect the abstract representer theorem with the explicit Green's function calculation from Section 1.

1. **Point evaluation is continuous.** By the Sobolev embedding $H_0^1(0,1) \hookrightarrow C^0[0,1]$, the map $f \mapsto f(s)$ is a bounded linear functional. The Riesz representation theorem therefore gives a unique element $K(\cdot, s) \in \mathcal{H}$ such that

   $$f(s) = \langle f,\, K(\cdot, s)\rangle_{\mathcal{H}} \qquad \forall f \in \mathcal{H}.$$

   To see what this means concretely, integrate by parts:

   $$\langle f,\, K(\cdot, s)\rangle = \int_0^1 f'(x)\, K_x(x,s)\, dx = -\int_0^1 f(x)\, K_{xx}(x,s)\, dx,$$

   where the boundary term vanishes because $f(0) = f(1) = 0$ and $K(\cdot, s) \in H_0^1(0,1)$ also vanishes at the endpoints. Thus the reproducing property $f(s) = \langle f,\, K(\cdot, s)\rangle$ holds exactly when

   $$-K_{xx}(x,s) = \delta(x-s) \quad \text{in the distributional sense.}$$

2. **This Riesz representer is the Green's function.** The function solving

   $$-\frac{d^2}{dx^2} K(x,s) = \delta(x-s), \qquad K(0,s) = K(1,s) = 0,$$

   is precisely the Green's function $G(x,s)$ computed earlier. Therefore

   $$K(x,s) = G(x,s)$$

   is the reproducing kernel of $H_0^1(0,1)$.

Now reconsider the interpolation problem

$$\min_{f\in\mathcal{H}} \Big\{ \mathcal{J}(f) : f(x_i) = y_i,\ i = 1,\dots,n \Big\}, \qquad \mathcal{J}(f) = \tfrac{1}{2}\|f\|_{\mathcal{H}}^2.$$

This setting satisfies the hypotheses of the representer theorem: the loss depends only on the point evaluations $f(x_i)$, and the penalty $\Phi(r) = \frac{1}{2}r^2$ is strictly increasing in the Hilbert norm. Hence every minimizer must lie in the span of the kernel sections:

$$\hat{f} \in \operatorname{span}\{\, K(\cdot, x_1), \dots, K(\cdot, x_n)\} = \operatorname{span}\{\, G(\cdot, x_1), \dots, G(\cdot, x_n)\}.$$

Thus the interpolant takes the explicit form

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i\, G(x, x_i),$$

which coincides with the solution obtained earlier through the Euler–Lagrange calculation. The PDE viewpoint identifies the kernel; the representer theorem explains why the minimizer must be a linear combination of its translates.

Conclusion. The model problem from Section 1 is not a special case: the representer theorem *forces* the interpolant in $H_0^1(0,1)$ to be a linear combination of Green's functions. The PDE derivation reveals what the kernel *is*, and the representer theorem explains *why this structure must occur*, connecting classical variational calculus directly with modern kernel methods.

3. Kernel Ridge Regression (KRR)

In practice one rarely demands exact interpolation. Noise, measurement error, overfitting, and model mismatch all motivate a regularized formulation. The natural relaxation of the minimum–norm problem (1) is the Tikhonov-regularized least-squares functional

$$\hat{f} = \arg\min_{g \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} |g(x_i) - y_i|^2 \; + \; \lambda \|g\|_{\mathcal{H}}^2 \right\}, \qquad \lambda > 0. \tag{2}$$

The first term encourages data fidelity; the second penalizes "energy" in the Hilbert norm of $\mathcal{H}$. As $\lambda \to 0$, one recovers interpolation; as λ grows, one prefers smoother functions.

Representer theorem structure

The objective in (2) depends on g only through the evaluations $g(x_i)$ and the norm $\|g\|_{\mathcal{H}}$. Since $\lambda \|g\|_{\mathcal{H}}^2$ is strictly increasing in $\|g\|_{\mathcal{H}}$, the Representer Theorem applies immediately. Therefore the minimizer *must* take the finite expansion

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i \, K(x, x_i),$$

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

Thus, although (2) is posed over an infinite-dimensional Hilbert space, the optimal solution always lives in the n-dimensional subspace spanned by the kernel sections.

Reduction to linear algebra

Substitute the expansion

$$g(x_j) = \sum_{i=1}^{n} \alpha_i K(x_j, x_i)$$

into the objective. Let the kernel matrix $K \in \mathbb{R}^{n \times n}$ be

$$K_{ij} = K(x_i, x_j).$$

Define $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ and $y = (y_1, \ldots, y_n)^T$. Then the data-fitting term becomes

$$\frac{1}{n} \sum_{j=1}^{n} \left| \sum_{i=1}^{n} \alpha_i K(x_j, x_i) - y_j \right|^2 = \frac{1}{n} \|K\alpha - y\|_2^2.$$

The Hilbert-space regularizer simplifies using the reproducing kernel property:

$$\|g\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) = \alpha^T K \, \alpha.$$

Therefore the infinite-dimensional problem (2) reduces to the finite-dimensional quadratic minimization

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{n} \|K\alpha - y\|_2^2 + \lambda \alpha^T K \alpha \right\}.$$

Setting the derivative to zero gives the normal equations

$$\frac{1}{n} K^T (K\alpha - y) + \lambda K \alpha = 0.$$

Since K is symmetric, this becomes

$$(K + n\lambda I)\, \alpha = y.$$

Thus the regression coefficients are

$$\alpha = (K + n\lambda I)^{-1} y,$$

and the predictor is

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i).$$

Interpretation: Where does the kernel come from?

KRR is *linear regression in a nonlinear feature space*. The feature map is not chosen manually but is determined entirely by the geometry of the Hilbert space $\mathcal{H}$:

$$x \;\longmapsto\; K(\cdot, x) \in \mathcal{H}.$$

How do we find K? The kernel is not guessed. No, it is *forced* upon us by the inner product on $\mathcal{H}$. By the Riesz representation theorem, continuity of point-evaluation implies that for each $s \in \Omega$ there exists a unique element $K(\cdot, s) \in \mathcal{H}$ satisfying

$$f(s) = \langle f,\, K(\cdot, s) \rangle_{\mathcal{H}} \qquad \forall f \in \mathcal{H}.$$

This K is the reproducing kernel.

Thus, *every choice of Hilbert norm generates its own kernel*. In Section 1 we took

$$\|f\|_{\mathcal{H}}^2 = \int_0^1 |f'(x)|^2 \, dx,$$

and the corresponding kernel turned out to be the Green's function of the operator $-d^2/dx^2$ with Dirichlet boundary conditions:

$$K(x, s) = G(x, s).$$

Had we chosen a different norm (e.g., H^2, Matérn spaces, weighted Sobolev spaces), the kernel would have changed accordingly. KRR never requires us to compute explicit feature maps, only the kernel determined by the geometry of $\mathcal{H}$.

Finite-dimensional collapse. Once K is known, the infinite-dimensional optimization reduces to

$$\hat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i), \qquad (K + n\lambda I)\alpha = y.$$

All "curvature" of the problem has been absorbed into the kernel matrix K.

Bias–variance tradeoff. The parameter λ controls smoothness:

- $\lambda \to 0$ recovers the minimum-norm interpolant of Section 1 (zero bias, high variance).
- Larger λ suppresses the RKHS norm (increased bias, reduced variance).

Examples of RKHSs and Their Kernels

1. Sobolev space $H_0^1(0,1)$. With inner product

$$\langle f,\, g\rangle_{H_0^1} = \int_0^1 f'(x)\, g'(x)\, dx,$$

the reproducing kernel is the Green's function of $-d^2/dx^2$ with Dirichlet boundary conditions:

$$K(x,s) = \begin{cases} x(1-s), & x \le s, \\ s(1-x), & x \ge s. \end{cases}$$

This is exactly the kernel that appeared in Section 1.

2. Gaussian RKHS on $\mathbb{R}$. Consider the Gaussian kernel

$$K(x,s) = \exp\!\left(-\frac{(x-s)^2}{2\sigma^2}\right), \qquad \sigma > 0.$$

This is the canonical example of a kernel that *does not* come from a differential operator. Instead it is governed by symmetry (translation-invariance) and spectral conditions.

Translation-invariance and Bochner's theorem. Since $K(x,s) = \kappa(x-s)$, all structure is encoded in the single function

$$\kappa(t) = \exp\!\left(-\tfrac{t^2}{2\sigma^2}\right).$$

Bochner's theorem gives the precise characterization: a function of the form $\kappa(x-s)$ is a positive-definite kernel if and only if κ is the Fourier transform of a finite nonnegative measure.

The Gaussian satisfies this perfectly:

$$\widehat{\kappa}(\omega) = \sqrt{2\pi}\,\sigma\, \exp\!\left(-\tfrac{\sigma^2\omega^2}{2}\right),$$

which is strictly positive and rapidly decaying. This positivity is what guarantees the existence of an RKHS, and its fast decay controls the smoothness of the resulting functions.

What the Gaussian RKHS actually is. Bochner's theorem also yields an explicit description of the Gaussian RKHS. A function f belongs to the space if and only if its Fourier transform is square-integrable against the reciprocal Gaussian weight:

$$\mathcal{H}_K = \left\{ f \in L^2_{\mathrm{loc}}(\mathbb{R}) : \int_{\mathbb{R}} \frac{|\widehat{f}(\omega)|^2}{\widehat{\kappa}(\omega)}\, d\omega < \infty \right\},$$

with inner product

$$\langle f,\, g\rangle_{\mathcal{H}_K} = \int_{\mathbb{R}} \frac{\widehat{f}(\omega)\,\overline{\widehat{g}(\omega)}}{\widehat{\kappa}(\omega)}\, d\omega.$$

Because

$$\frac{1}{\widehat{\kappa}(\omega)} \sim \exp\Big(\tfrac{\sigma^2\omega^2}{2}\Big),$$

membership in this RKHS forces the Fourier transform of f to decay *faster than Gaussian*. This is an extremely strong condition.

Consequences: an ultra-smooth function space. Every function in the Gaussian RKHS:

- extends to an entire function on $\mathbb{C}$ (real-analytic with global control),
- has all derivatives bounded by Gaussian envelopes,
- and fluctuates only on scales determined by σ.

In short: Gaussian kernels generate *the smoothest possible* RKHS compatible with translation invariance.

3. Matérn RKHS on $\mathbb{R}^d$. A third fundamental family of RKHS kernels is provided by the Matérn class. For parameters $\nu > 0$ (smoothness) and $\ell > 0$ (length scale), the Matérn kernel is

$$K_{\nu,\ell}(x,s) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} \|x-s\| \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} \|x-s\| \right),$$

where K_ν is the modified Bessel function of the second kind.

This kernel is translation-invariant:

$$K_{\nu,\ell}(x,s) = \kappa_{\nu,\ell}(x-s), \qquad \kappa_{\nu,\ell}(t) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} \|t\| \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}}{\ell} \|t\| \right).$$

Fourier characterization. Bochner's theorem again gives the full structure. The Matérn kernel is positive definite because its Fourier transform is the strictly positive spectral density

$$\widehat{\kappa}_{\nu,\ell}(\omega) \propto \left(\frac{2\nu}{\ell^2} + \|\omega\|^2 \right)^{-(\nu+d/2)}.$$

This identifies the Matérn RKHS as the space

$$\mathcal{H}_{K_{\nu,\ell}} = \left\{ f \in L^2_{\mathrm{loc}}(\mathbb{R}^d) : \int_{\mathbb{R}^d} |\widehat{f}(\omega)|^2 \left(\frac{2\nu}{\ell^2} + \|\omega\|^2 \right)^{\nu+d/2} d\omega < \infty \right\},$$

with inner product

$$\langle f,\, g \rangle_{\mathcal{H}_{K_{\nu,\ell}}} = \int_{\mathbb{R}^d} \widehat{f}(\omega)\, \overline{\widehat{g}(\omega)} \left(\frac{2\nu}{\ell^2} + \|\omega\|^2 \right)^{\nu+d/2} d\omega.$$

Interpretation. Because the multiplier grows like $(1 + \|\omega\|^2)^{\nu+d/2}$, functions in the Matérn RKHS have Sobolev smoothness:

$$\mathcal{H}_{K_{\nu,\ell}} \text{ is norm-equivalent to } H^{\nu+d/2}(\mathbb{R}^d).$$

Thus the parameter ν directly controls differentiability:

- $\nu = \frac{1}{2}$ gives the exponential kernel (nondifferentiable functions),

- larger ν gives smoother functions,
- $\nu \to \infty$ yields the Gaussian kernel with real-analytic functions.

The Matérn class is therefore crucial in applications: it provides RKHSs with *tunable* Sobolev regularity, aligning naturally with the smoothness of many PDE solution maps.

What about L^2??? Although $L^2(0,1)$ is a Hilbert space, it does *not* admit a reproducing kernel. The reason is simple: point evaluation is not continuous in L^2. A function in L^2 is an equivalence class (defined up to sets of measure zero), so the value $f(s)$ is not even well defined. More importantly, one can construct a sequence $\{f_k\}$ with $\|f_k\|_{L^2} = 1$ but $|f_k(s)| \to \infty$ for any fixed s, showing that the map $f \mapsto f(s)$ cannot be a bounded linear functional (left as an exercise to the reader).

By the Riesz representation theorem, a reproducing kernel K would satisfy

$$f(s) = \langle f,\, K(\cdot, s)\rangle_{L^2},$$

but this identity would require $K(\cdot, s)$ to be a Dirac delta, which lies *outside* L^2.

Thus:

- L^2 is not an RKHS,
- no reproducing kernel exists,
- Kernel Ridge Regression cannot be formulated in L^2 without changing the function space.

4. From Function Learning to Operator Learning

So far we have learned a *function*

$$f : \Omega \to \mathbb{R}$$

from pointwise data. In operator learning, the goal is to learn a map between function spaces

$$\mathcal{G}^\dagger : \mathcal{U} \to \mathcal{V},$$

where $\mathcal{U}$ and $\mathcal{V}$ are typically infinite-dimensional (e.g. Sobolev spaces of input and output fields for a PDE).

In applications one never observes full functions. What is available are finite collections of *linear measurements* of the input and output.

4.1 Measurement Model

Let

$$\phi : \mathcal{U} \to \mathbb{R}^n, \qquad \varphi : \mathcal{V} \to \mathbb{R}^m$$

be bounded linear measurement operators. Typical examples include:

- point evaluations at sensor locations,
- local averages over grid cells,
- discrete Fourier or wavelet coefficients.

Given $u_0 \in \mathcal{U}$, we never see $\mathcal{G}^\dagger(u_0)$ directly. We only see

$$x = \phi(u_0) \in \mathbb{R}^n, \qquad y = \varphi(\mathcal{G}^\dagger(u_0)) \in \mathbb{R}^m.$$

From data alone, the operator-learning task is therefore:

$$\text{learn the finite-dimensional map} \quad x = \phi(u_0) \;\longmapsto\; y = \varphi(\mathcal{G}^\dagger(u_0)).$$

To connect this back to functions we introduce reconstruction maps.

4.2 Reconstruction Maps and Optimal Recovery

We want maps

$$\psi : \mathbb{R}^n \to \mathcal{U}, \qquad \chi : \mathbb{R}^m \to \mathcal{V}$$

that "lift" measurements back into function space in a principled way.

Optimal recovery viewpoint. Given a measurement $x = \phi(u)$, many $u \in \mathcal{U}$ produce the same x. Optimal recovery chooses, among all such u, the one of *minimal* norm:

$$\psi(x) = \arg\min_{u \in \mathcal{U}} \big\{ \|u\|_{\mathcal{U}} : \phi(u) = x \big\}.$$

This is the infinite-dimensional analogue of the minimum-norm solution to an underdetermined linear system. It removes any unnecessary oscillations or components invisible to ϕ.

Similarly, on the output side,

$$\chi(y) = \arg\min_{v \in \mathcal{V}} \big\{ \|v\|_{\mathcal{V}} : \varphi(v) = y \big\}.$$

How ψ and χ are computed in practice. Numerically we work in a finite basis:

$$u(\cdot) \;\leftrightarrow\; \mathbf{u} \in \mathbb{R}^N, \qquad \phi \;\leftrightarrow\; A \in \mathbb{R}^{n \times N},$$

so that $x = \phi(u)$ becomes $x = A\mathbf{u}$.

The optimal-recovery lift is then the minimum-norm solution of

$$A\mathbf{u} = x : \qquad \mathbf{u}^\star = A^\top (AA^\top)^{-1} x = A^+ x,$$

the Moore–Penrose pseudoinverse. In practice A^+ is computed via QR or SVD.

The same construction with a matrix B for φ gives

$$\mathbf{v}^\star = B^+ y \quad \Longleftrightarrow \quad \chi(y).$$

Thus, after discretization,

$$\psi(x) \approx A^+ x, \qquad \chi(y) \approx B^+ y.$$

The norm on $\mathcal{U}$ (Sobolev, Gaussian RKHS, Matérn RKHS, etc.) determines what "minimal" means, like energy in derivatives, weighted Fourier energy, Sobolev energy, and so on.

4.3 Reduction to Vector-Valued Regression

Once $\phi, \psi, \varphi, \chi$ are fixed, the infinite-dimensional operator $\mathcal{G}^\dagger$ induces a finite-dimensional map

$$f^\dagger := \varphi \circ \mathcal{G}^\dagger \circ \psi : \mathbb{R}^n \to \mathbb{R}^m.$$

All training pairs have the form

$$x_i = \phi(u_i), \qquad y_i = \varphi(\mathcal{G}^\dagger(u_i)), \qquad i = 1, \dots, N,$$

so learning $\mathcal{G}^\dagger$ from data is equivalent to learning the vector-valued function $f^\dagger$.

Vector-valued RKHS and representer theorem. To learn $f^\dagger$, we place it in a vector-valued RKHS $\mathcal{H}_K$ defined by a positive-definite matrix-valued kernel

$$K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{m \times m}.$$

The vector-valued representer theorem says that the KRR minimizer has the form

$$\hat{f}(x) = \sum_{i=1}^{N} K(x, x_i)\, \alpha_i, \qquad \alpha_i \in \mathbb{R}^m,$$

where $\{\alpha_i\}$ are obtained by solving the regularized problem

$$\hat{f} = \arg\min_{g \in \mathcal{H}_K} \left\{ \frac{1}{N} \sum_{i=1}^{N} \|g(x_i) - y_i\|_{\mathbb{R}^m}^2 + \lambda \|g\|_{\mathcal{H}_K}^2 \right\}.$$

If $K(x, x') = k(x, x')\, I_m$ is separable, this reduces to m decoupled scalar KRR problems with the same scalar kernel k.

Reassembling the operator. With $\hat{f}$ in hand, the learned surrogate operator is

$$\boxed{\bar{\mathcal{G}} = \chi \circ \hat{f} \circ \phi.}$$

In words:

$$u \;\xmapsto{\phi}\; x \in \mathbb{R}^n \;\xmapsto{\hat{f}}\; y \in \mathbb{R}^m \;\xmapsto{\chi}\; \bar{\mathcal{G}}(u) \in \mathcal{V}.$$

The first and last arrows are optimal-recovery lifts (minimum-norm reconstructions); the middle arrow is kernel ridge regression.

4.4 Summary

- Operator learning is recast as learning a finite-dimensional map $f^\dagger : \mathbb{R}^n \to \mathbb{R}^m$ induced by measurements.
- Optimal recovery provides canonical reconstruction maps ψ, χ by minimum-norm lifting (pseudoinverse after discretization).
- Vector-valued KRR with kernel K learns $f^\dagger$ in an RKHS that encodes the regularity of $\mathcal{G}^\dagger$.
- The final surrogate is $\bar{\mathcal{G}} = \chi \circ \hat{f} \circ \phi$.

This is exactly the structural reduction exploited in *Kernel Methods are Competitive for Operator Learning*.

5. How the Batlle–Darcy–Hosseini–Owhadi Framework Fits In

By this point we have:

- minimum-norm interpolation and Green's functions,
- RKHSs and the representer theorem,
- kernel ridge regression as regularized interpolation,
- the reduction from operator learning to vector-valued regression via measurement and reconstruction maps $(\phi, \psi, \varphi, \chi)$.

The paper *Kernel Methods are Competitive for Operator Learning* takes exactly this machinery and pushes it in two directions:

(a) a clean *information-theoretic* formulation of operator learning as optimal recovery, and

(b) a careful comparison of kernel methods with neural operators in that setting.

5.1 The architecture in one line

In the notation of Section 4, the learned surrogate for the unknown operator $\mathcal{G}^\dagger : \mathcal{U} \to \mathcal{V}$ has the form

$$\bar{\mathcal{G}} = \chi \circ \hat{f} \circ \phi,$$

where:

- ϕ and φ are fixed measurement operators,
- ψ and χ are fixed optimal-recovery lifts (minimum-norm reconstructions),
- $\hat{f} : \mathbb{R}^n \to \mathbb{R}^m$ is obtained by vector-valued KRR in an RKHS chosen to encode the regularity of $\mathcal{G}^\dagger$.

Conceptually:

$$\text{function} \xrightarrow{\phi} \mathbb{R}^n \xrightarrow{\hat{f}} \mathbb{R}^m \xrightarrow{\chi} \text{function}.$$

All the nonlinearity and learning happen in the middle map $\hat{f}$; everything else is geometry and optimal recovery.

5.2 What the paper actually proves (in words)

Very roughly, the framework shows:

- For a given information channel (ϕ, φ) and prior regularity encoded by an RKHS, the optimal-recovery lifts (ψ, χ) and KRR estimator $\hat{f}$ are *minimax optimal* among all methods that use the same measurements. No other estimator can uniformly beat their worst-case error by more than a constant factor.
- The error rates are governed by notions like effective dimension and n-widths of the image of $\mathcal{G}^\dagger$ under the measurement operators. This replaces "network capacity" heuristics with sharp approximation-theoretic quantities.

- In numerical experiments (elliptic and parabolic PDEs, etc.), kernel surrogates built this way perform at the same level as Fourier neural operators and DeepONets, often with fewer samples and much cheaper training (convex optimization instead of nonconvex SGD).

The take-home message is not that "KRR beats neural nets," but that *once you respect the measurement geometry and use optimal recovery*, classical kernel methods are fully competitive with neural operators on many benchmark operator-learning tasks. So the paper should be read less as "a new algorithm" and more as a clarification: much of operator learning can be understood as RKHS theory + optimal recovery + linear algebra.