Learning from Non-Random Data in Hilbert Spaces: An Optimal Recovery Perspective

Simon Foucart[*], Chunyang Liao[*], Shahin Shahrampour[†], and Yinsong Wang[†]

Abstract

The notion of generalization in classical Statistical Learning is often attached to the postulate that data points are independent and identically distributed (IID) random variables. While relevant in many applications, this postulate may not hold in general, encouraging the development of learning frameworks that are robust to non-IID data. In this work, we consider the regression problem from an Optimal Recovery perspective. Relying on a model assumption comparable to choosing a hypothesis class, a learner aims at minimizing the worst-case error, without recourse to any probabilistic assumption on the data. We first develop a semidefinite program for calculating the worst-case error of any recovery map in finite-dimensional Hilbert spaces. Then, for any Hilbert space, we show that Optimal Recovery provides a formula which is user-friendly from an algorithmic point-of-view, as long as the hypothesis class is linear. Interestingly, this formula coincides with kernel ridgeless regression in some cases, proving that minimizing the average error and worst-case error can yield the same solution. We provide numerical experiments in support of our theoretical findings.

Key words and phrases: Optimal Recovery, approximability models, worst-case errors, Hilbert spaces.

1 Introduction

Let us place ourselves in a classical scenario where data about an unknown function f_0 take the form

$$y_i = f_0(\mathbf{x}_i), \qquad i \in [1:m]. \tag{1}$$

The values $y_i \in \mathbb{R}$ and the evaluations points $\mathbf{x}_i \in \Omega \subseteq \mathbb{R}^d$ are available to the learner. The goal is to 'learn' the function f_0 from the data (1) by producing a surrogate function $\hat{f}$ for f_0.

[*] Simon Foucart and Chunyang Liao are with the Department of Mathematics, Texas A&M University, College Station, TX 77843 USA.

[†] Shahin Shahrampour and Yinsong Wang are with the Department of Mechanical & Industrial Engineering at Northeastern University. This work was initiated when they were both with the Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX 77843 USA.

Supervised Machine Learning methods compute such an $\hat{f}$ from a hypothesis class selected in advance. The performance of a method then depends on the choice of this hypothesis class: a good class should obviously approximate functions of interest well. This translates into a small *approximation error*, which is one of the constituents towards the total error of a method. Another constituent is the *estimation error*. In classical Statistical Learning [27], the latter is often analyzed by adopting a postulate that the $\mathbf{x}_i$'s are independent realizations of a random variable with an unknown distribution on Ω. While relevant in many applications, this postulate may not hold in general, encouraging the development of learning frameworks that are robust to non-IID data. In this work, we consider the regression problem from an Optimal Recovery perspective, without recourse to any probabilistic assumption on the data. Indeed, in the absence of randomness, an average-case analysis is not possible anymore. Instead, the learner aims at minimizing the worst-case (prediction) error by relying on a model assumption comparable to choosing a hypothesis class. We restrict our attention here to Hilbert spaces and provide the following contributions:

- We develop a numerical framework for calculating the worst-case error in the case of finite-dimensional Hilbert spaces. In particular, we show that this error can be computed via a semidefinite program (Theorem 1).
- We show that Optimal Recovery provides a formula which is user-friendly from an algorithmic point-of-view when the hypothesis class is a linear subspace (Theorem 2). Interestingly, this formula coincides with kernel ridgeless regression in some cases (Theorem 3), proving that minimizing the average error and worst-case error can yield the same solution.

The theoretical findings are verified through some numerical experiments presented in Section 5.

1.1 Why Optimal Recovery?

The theory of Optimal Recovery was developed in the 70's-80's as a subfield of Approximation Theory (see the surveys [18, 19]). Its development was shaped by concurrent developments in the theory of spline functions (see e.g. [5, 10]). Splines provided a rare example where the theory integrated computations [6]. But, at that time, algorithmic issues were not the high priority that they have become today and theoretical questions such as the existence of linear optimal algorithms prevailed (see e.g. the survey [22]). Arguably, this neglect hindered the development of the topic and this work can be seen as an attempt to promote an algorithmic framework that sheds light on similarities and differences between Optimal Recovery (in Hilbert spaces) and Statistical Learning. Incidentally, what is sometimes called the *spline algorithm* in Optimal Recovery has recently made a reappearance in Machine Learning circles as minimum-norm interpolation [2, 25, 17], of course with a different motivation. Optimal Recovery also establishes a correspondence between numerical Approximation Theory and Gaussian process regression from a game-theoretic point-of-view, see e.g. [21]. We remark that Optimal Recovery is not the only framework dealing with non-IID data. There are indeed other strands of Machine Learning literature (e.g. Online Learning [15] and

Federated Learning [29]) that investigate learning from non-IID and/or non-random data. But, to be clear, Optimal Recovery does not rely on any statistical assumptions and aims at worst-case guarantees rather than average-case guarantees.

1.2 Noisy observations

A careful reader may wonder about the possibility of incorporating an error $e_i \in \mathbb{R}$ in the data $y_i = f_0(\mathbf{x}_i) + e_i$, which is a common consideration in Machine Learning. We do not investigate such a scenario in this work, as our main focus is on drawing interesting connections between Optimal Recovery and some of the common Supervised Learning techniques in the simplest of settings first. Future works[1] will concentrate on this inaccurate scenario which, despite some existing results (see [23, 11, 1]), presents some unsuspected subtleties. For instance, the results from [1] are only valid in the complex setting and not in the real setting considered here.

2 The Optimal Recovery Perspective

In this section, we recall the general framework of Optimal Recovery and highlight some novel results, including the computation of worst-case error and the explicit formula of optimal recovery map.

2.1 The function space

Echoing the theory of Optimal Recovery, we consider the function f_0 more abstractly as an element from a normed space $\mathcal{F}$. The output data y_i's, which are evaluations of f_0 at the points $\mathbf{x}_i$'s, can be generalized to linear functionals ℓ_i's applied to f_0, so that the data take the form

$$y_i = \ell_i(f_0), \qquad i \in [1:m]. \tag{2}$$

For convenience, we summarize these data as

$$\mathbf{y} = L(f_0) = [\ell_1(f_0); \ldots; \ell_m(f_0)] \in \mathbb{R}^m, \tag{3}$$

where the linear map $L : \mathcal{F} \to \mathbb{R}^m$ is called the observation operator. Relevant situations include the case where $\mathcal{F}$ is the space $\mathcal{C}(\Omega)$ of continuous functions on Ω, which is equipped with the uniform norm, and the case where $\mathcal{F}$ is a Hilbert space $\mathcal{H}$, which is equipped with the norm derived from its inner product. It is the latter case that is the focus of this work. More precisely, after

[1] At the publication time of the current article, a portion of these works is already available, see [13].

recalling some known results, we concentrate on a reproducing kernel Hilbert space $\mathcal{H}$ of functions defined on Ω, so that the point evaluations at the $\mathbf{x}_i$'s are indeed well-defined and continuous linear functionals on $\mathcal{H}$.

2.2 The model set

Without further information, data by themselves are not sufficient to say anything meaningful about f_0. For example, one could think of all ways to fit a univariate function through points $(x_1, y_1), \ldots, (x_m, y_m) \in \mathbb{R}^2$ if no restriction is imposed. Thus, a model assumption for the functions of interest is needed. This assumption takes the form

$$f_0 \in \mathcal{K}, \tag{4}$$

where the model set $\mathcal{K}$ translates an educated belief about the behavior of realistic functions f_0. In Optimal Recovery, the set $\mathcal{K}$ is often chosen to be a convex and symmetric subset of $\mathcal{F}$. Here, our relevant modeling assumption is the one that occurs implicitly in Machine Learning, namely that the functions of interest are well-approximated by suitable hypothesis classes. In this work, we only consider hypothesis classes that are linear subspaces V of $\mathcal{F}$. Thus, given an approximation parameter $\epsilon > 0$ (the targeted approximation error), our model set has the form

$$\mathcal{K} := \{f \in \mathcal{F} : \mathrm{dist}(f, V) \leq \epsilon\}, \tag{5}$$

where $\mathrm{dist}(f, V) := \inf\{\|f - v\|_{\mathcal{F}}, v \in V\}$. In the case $\mathcal{F} = \mathcal{H}$ of a Hilbert space, this model set reads

$$\mathcal{K} = \{f \in \mathcal{H} : \|f - P_V f\|_{\mathcal{H}} \leq \epsilon\}, \tag{6}$$

where $P_V f$ is the orthogonal projection of f onto the subspace V. Such an approximability set was put forward in [3], with motivation coming from parametric PDEs. When working with this model, it is implicitly assumed that

$$V \cap \ker(L) = \{0\}, \tag{7}$$

otherwise the existence of a nonzero $v \in V \cap \ker(L)$ would imply that each $f_t := f_0 + tv$, $t \in \mathbb{R}$, is both data-consistent ($L(f_t) = \mathbf{y}$) and model-consistent ($f_t \in \mathcal{K}$), leading to infinite worst-case error by letting $t \to \infty$. By a dimension argument, the assumption (7) forces

$$n := \dim(V) \leq m, \tag{8}$$

i.e., we must place ourselves in an underparametrized regime where there are less model parameters than datapoints. To make sense of the overparametrized regime, the model set (5) would need to be refined by adding some boundedness conditions, see [12] for results in this direction.

2.3 Worst-case errors

With the model set in place, we now need to assess the performance of a learning/recovery map, which is just a map taking data $\mathbf{y} \in \mathbb{R}^m$ as input and returning an element $\hat{f} \in \mathcal{F}$ as output. Given a model set $\mathcal{K}$, the local worst-case error of such a map $R : \mathbb{R}^m \to \mathcal{F}$ at $\mathbf{y} \in \mathbb{R}^m$ is

$$\mathrm{err}_{\mathcal{K}}^{\mathrm{loc}}(L, R(\mathbf{y})) := \sup_{f \in \mathcal{K}, L(f) = \mathbf{y}} \|f - R(\mathbf{y})\|_{\mathcal{F}}. \tag{9}$$

The global worst-case error is the worst local worst-case error over all $\mathbf{y} \in \mathbb{R}^m$ that can be obtained by observing some $f \in \mathcal{K}$, i.e.,

$$\mathrm{err}_{\mathcal{K}}^{\mathrm{glo}}(L, R) := \sup_{f \in \mathcal{K}} \|f - R(L(f))\|_{\mathcal{F}}. \tag{10}$$

A learning/recovery map $R : \mathbb{R}^m \to \mathcal{F}$ is called locally, respectively globally, optimal if it minimizes the local, respectively global, worst-case error. These definitions can be extended to handle not only the full recovery of f_0 but also the recovery of a quantity of interest $Q(f_0)$. That is, for a map $Q : \mathcal{F} \to Z$ from $\mathcal{F}$ into another normed space Z, one would define e.g. the global worst-case error of the learning/recovery map $R : \mathbb{R}^m \to Z$ as

$$\mathrm{err}_{\mathcal{K},Q}^{\mathrm{glo}}(L, R) := \sup_{f \in \mathcal{K}} \|Q(f) - R(L(f))\|_{Z}. \tag{11}$$

Such a framework is pertinent even if we target the full recovery of f_0 but with performance evaluated in a norm $/\!\!/ \cdot /\!\!/_{\mathcal{F}}$ different from the native norm $\|\cdot\|_{\mathcal{F}}$, as we can consider Q to be the identity map from $\mathcal{F}$ equipped with $\|\cdot\|_{\mathcal{F}}$ into $Z = \mathcal{F}$ equipped with $/\!\!/ \cdot /\!\!/_{\mathcal{F}}$.

Perhaps counterintuitively, dealing with the global setting is somewhat easier than dealing with the local setting, in the sense that globally optimal maps have been obtained in situations where locally optimal maps have not, e.g. when $\mathcal{F} = \mathcal{C}(\Omega)$. Accordingly, it is the local setting which is the focus of this work.

2.4 Computation of local worst-case errors

When $\mathcal{F} = \mathcal{H}$ is a Hilbert space and the approximability model (6) is selected, determining the local worst-case error of a given map $R : \mathbb{R}^m \to \mathcal{H}$ at some $\mathbf{y}$ involves solving

$$\underset{f \in \mathcal{H}}{\mathrm{maximize}} \, \|f - R(\mathbf{y})\|_{\mathcal{H}} \quad \text{s.to} \quad \begin{cases} \|f - P_V f\|_{\mathcal{H}} \le \epsilon, \\ L(f) = \mathbf{y}. \end{cases} \tag{12}$$

This is a nonconvex optimization program, and as such does appear hard to solve at first sight. However, it is a quadratically constrained quadratic program, hence it is possible to solve it exactly.

Although Gurobi [14] now features direct capabilities to solve quadratically constrained quadratic programs, we take the route of recasting (12) as a semidefinite program using the S-lemma [24]. The solution of the recast program can then be obtained using an off-the-shelf semidefinite solver, at least when the dimension $N = \dim(\mathcal{H})$ of the Hilbert $\mathcal{H}$ space is finite. Precisely, with $(h_1, \ldots, h_N)$ denoting an orthonormal basis for $\mathcal{H}$ chosen in such a way that $(h_1, \ldots, h_{N-m})$ is an orthonormal basis for $\ker(L)$ and with H denoting the unitary map $x \in \mathbb{R}^{N-m} \mapsto \sum_{k=1}^{N-m} x_k h_k \in \ker(L)$, local worst-case errors can be computed based on the following observation.

Theorem 1. The local worst-case error of a learning/recovery map $R : \mathbb{R}^m \to \mathcal{H}$ at $\mathbf{y} \in \mathbb{R}^m$ under the model set (6) can be expressed, with $g := R(\mathbf{y})$, as

$$e_{\mathcal{K}}^{\mathrm{loc}}(L, g) = \Big[\|h - P_{\ker(L)^\perp}(g)\|_{\mathcal{H}}^2 + \|P_{\ker(L)}(g)\|_{\mathcal{H}}^2 + c^\star \Big]^{1/2}, \tag{13}$$

where h is the unique element in $\ker(L)^\perp$ satisfying $L(h) = \mathbf{y}$ and $c^\star$ is the minimal value of the following program, in which $w := P_{V^\perp}(h)$:

$$\underset{c,d \in \mathbb{R}}{\text{minimize}} \quad c \quad \text{s.to} \quad d \geq 0 \quad \text{and} \tag{14}$$
$$\left[\begin{array}{c|c} H^*(dP_{V^\perp} - I_{\mathcal{H}})H & H^*(dw + P_{\ker(L)}(g)) \\ \hline (dw + P_{\ker(L)}(g))^* H & c + d(\|w\|_{\mathcal{H}}^2 - \epsilon^2) \end{array} \right] \succeq 0.$$

Proof. We first justify that there is a unique $h \in \ker(L)^\perp$ such that $L(h) = \mathbf{y} \in \mathbb{R}^m$. To see this, define the linear map $\tilde{L} : h \in \ker(L)^\perp \mapsto L(h) \in \mathrm{range}(L)$. Since $\ker(\tilde{L}) = \ker(L) \cap \ker(L)^\perp = \{0\}$, the map $\tilde{L}$ must be injective. Therefore, we have $\dim(\mathrm{range}(\tilde{L})) = \dim(\ker(L)^\perp)$, which equals $N - \dim(\ker(L)) = \dim(\mathrm{range}(L))$ by the rank-nullity theorem, so the map $\tilde{L}$ is also surjective. Thus, the claim is justified by the fact that $\tilde{L}$ is bijective.

Next, the squared local worst-case error (9) at $g = R(\mathbf{y})$ is

$$\big[\mathrm{err}_{\mathcal{K}}^{\mathrm{loc}}(L, g)\big]^2 = \sup_{f \in \mathcal{H}} \big\{ \|f - g\|_{\mathcal{H}}^2 : \|P_{V^\perp} f\|_{\mathcal{H}}^2 \leq \epsilon^2, L(f) = \mathbf{y} \big\}. \tag{15}$$

Decomposing f and g as $f = f' + f''$ and $g = g' + g''$ with $f', g' \in \ker(L)$ and $f'', g'' \in \ker(L)^\perp$, the condition $L(f) = \mathbf{y}$ reduces to $L(f'') = \mathbf{y}$, i.e., $f'' = h$ is uniquely determined. The condition $\|P_{V^\perp} f\|_{\mathcal{H}}^2 \leq \epsilon^2$ then becomes $\|P_{V^\perp} f' + w\|_{\mathcal{H}}^2 \leq \epsilon^2$. As for the expression to maximize, it separates into

$$\|f - g\|_{\mathcal{H}}^2 = \|f'' - g''\|_{\mathcal{H}}^2 + \|f' - g'\|_{\mathcal{H}}^2 = \|h - g''\|_{\mathcal{H}}^2 + \|g'\|_{\mathcal{H}}^2 + \|f'\|_{\mathcal{H}}^2 - 2\langle f', g' \rangle. \tag{16}$$

Up to the additive constant $\|h - g''\|_{\mathcal{H}}^2 + \|g'\|_{\mathcal{H}}^2$, the maximum in (15) is now

$$\begin{aligned} &\sup_{f' \in \ker(L)} \|f'\|_{\mathcal{H}}^2 - 2\langle f', g' \rangle \quad \text{s.to } \|P_{V^\perp} f' + w\|_{\mathcal{H}}^2 \leq \epsilon^2 \\ &= \inf_{c \in \mathbb{R}} c \quad \text{s.to } \|f'\|_{\mathcal{H}}^2 - 2\langle f', g' \rangle \leq c \text{ when } \|P_{V^\perp} f' + w\|_{\mathcal{H}}^2 \leq \epsilon^2. \end{aligned} \tag{17}$$

Writing $f' = Hx$ with $x \in \mathbb{R}^{N-m}$, this latter constraint reads

$$\begin{aligned} & c - \big(\langle Hx, Hx\rangle - 2\langle Hx, g'\rangle\big) \geq 0 \\ (18) \qquad & \text{whenever } \epsilon^2 - \big(\langle P_{V^\perp} Hx, P_{V^\perp} Hx\rangle + 2\langle P_{V^\perp} Hx, w\rangle + \|w\|_{\mathcal{H}}^2\big) \geq 0. \end{aligned}$$

By the S-lemma, see e.g. [24], (18) is equivalent to the existence of $d \geq 0$ such that

$$(19) \qquad c - \big(\langle Hx, Hx\rangle - 2\langle Hx, g'\rangle\big) \geq d\left[\epsilon^2 - \big(\langle P_{V^\perp} Hx, P_{V^\perp} Hx\rangle + 2\langle P_{V^\perp} Hx, w\rangle + \|w\|_{\mathcal{H}}^2\big)\right]$$

for all $x \in \mathbb{R}^{N-m}$, or in other words, to the existence of $d \geq 0$ such that

$$(20) \qquad \big(d\langle x, (H^* P_{V^\perp} H)x\rangle - \langle x, H^* H x\rangle\big) + 2\big(d\langle x, H^* w\rangle + \langle x, H^* g'\rangle\big) + c + d(\|w\|_{\mathcal{H}}^2 - \epsilon^2) \geq 0$$

for all $x \in \mathbb{R}^{N-m}$. This constraint can be reformulated as a semidefinite constraint

$$(21) \qquad \left[\begin{array}{c|c} dH^* P_{V^\perp} H - H^* H & dH^* w + H^* g' \\ \hline (dH^* w + H^* g')^* & c + d(\|w\|_{\mathcal{H}}^2 - \epsilon^2) \end{array}\right] \succeq 0.$$

Keeping in mind that $g' = P_{\ker(L)} g$, this is the semidefinite constraint appearing in (14). Putting everything together, we arrive at the expression for the local worst-case error announced in (13). $\square$

2.5 Optimal learning/recovery map

Even though it is possible to compute the minimal worst-case error via (13)-(14), optimizing over $g \in \mathcal{H}$ to produce the locally optimal recovery map would still require some work and would in fact be a major overkill. Indeed, for our situation of interest, some crucial work in this direction has been carried out in [3], and we rely on it to derive the announced user-friendly formula for the locally (hence globally, too) optimal recovery map $R^{\rm opt}$. Precisely, when $\mathcal{F} = \mathcal{H}$ is a (finite- or infinite-dimensional) Hilbert space and the model set $\mathcal{K}$ is given by (6), it was shown in [3] that, for any input $\mathbf{y} \in \mathbb{R}^m$, the output $R^{\rm opt}(\mathbf{y}) \in \mathcal{H}$ is the solution $\hat{f}$ to the convex minimization program

$$(22) \qquad \underset{f \in \mathcal{H}}{\text{minimize}}\ \|f - P_V f\|_{\mathcal{H}} \qquad \text{subject to } L(f) = \mathbf{y}.$$

We generalize this result through Theorem 4 in the appendix. It suffices to say for now that the argument of [3], based on the original expression (9) of the worst-case error, exploits the fact that $\hat{f} - P_V \hat{f}$ is orthogonal not only to V but also to $\ker(L)$. Let us point out that $R^{\rm opt}(\mathbf{y}) = \hat{f}$ is both data-consistent and model-consistent when $\mathbf{y} = L(f_0)$ for some $f_0 \in \mathcal{K}$. It is also interesting to note that the optimal recovery map $R^{\rm opt}$ does not depend on the approximation parameter ϵ. This peculiarity disappears as soon as observation errors are taken into consideration, see [11].

A computable expression for the minimal local error (9), and in turn for the minimal global error (10), has also been given in [3]. Without going into details, we only want to mention that the latter decouples as the product $\mu \times \epsilon$ of an indicator μ of compatibility between model and datapoints, which increases as the space V is enlarged, and of the parameter ϵ of approximability, which decreases as the space V is enlarged. Thus, the choice of a space V yielding small minimal worst-case errors involves a trade-off on $n = \dim(V)$. This trade-off is illustrated numerically in Subsection 5.2. Alternatively, viewing the space V as fixed, if we could choose the observation functionals ℓ_i's (which is not the focus here), then the compatibility indicator μ would reflect the quality of these functionals.

Although the description given by of the optimal learning/recovery map is quite informative, it fails to make apparent the fact the map $R^{\rm opt}$ is actually a linear map. This fact can be seen from the theorem below, which states that solving a minimization program for each $\mathbf{y} \in \mathbb{R}^m$ is not needed to produce $R^{\rm opt}(\mathbf{y})$. Indeed, one can obtain $R^{\rm opt}(\mathbf{y})$ by some linear algebra computations involving two matrices which are more or less directly available to the learner. To define these matrices, we need the Riesz representers $u_i \in \mathcal{H}$ of the linear functionals $\ell_i \in \mathcal{H}^*$, which are characterized by

$$\ell_i(f) = \langle u_i, f \rangle \qquad \text{for all } f \in \mathcal{H}.$$

We also need a (not necessarily orthonormal) basis $(v_1, \ldots, v_n)$ for V. The two matrices are the Gramian $\mathbf{G} \in \mathbb{R}^{m \times m}$ of $(u_1, \ldots, u_m)$ and the cross-Gramian $\mathbf{C} \in \mathbb{R}^{m \times n}$ of $(u_1, \ldots, u_m)$ and $(v_1, \ldots, v_n)$. Their entries are given, for $i, i' \in [1:m]$ and $j \in [1:n]$, by

$$(\mathbf{G})_{i,i'} = \langle u_i, u_{i'} \rangle = \ell_i(u_{i'}), \tag{23}$$

$$\mathbf{C}_{i,j} = \langle u_i, v_j \rangle = \ell_i(v_j). \tag{24}$$

The matrix $\mathbf{G}$ is positive definite and in particular invertible (linear independence of the ℓ_i's is assumed). The matrix $\mathbf{C}$ has full rank thanks to the assumption $V \cap \ker(L) = \{0\}$. The result below shows that the output of the optimal learning/recovery map does not have to lie in the space V (the hypothesis class), as opposed to the output of algorithms such as empirical risk minimizations.

Theorem 2. The locally optimal learning/recovery map $R^{\rm opt} : \mathbb{R}^m \to \mathcal{H}$ is given in closed form for each $\mathbf{y} \in \mathbb{R}^m$ by

$$R^{\rm opt}(\mathbf{y}) = \sum_{i=1}^m a_i u_i + \sum_{j=1}^n b_j v_j, \tag{25}$$

where the coefficient vectors $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b} \in \mathbb{R}^n$ are computed as

$$\mathbf{b} = (\mathbf{C}^\top \mathbf{G}^{-1} \mathbf{C})^{-1} \mathbf{C}^\top \mathbf{G}^{-1} \mathbf{y}, \tag{26}$$

$$\mathbf{a} = \mathbf{G}^{-1}(\mathbf{y} - \mathbf{C}\mathbf{b}). \tag{27}$$

Proof. Let $\hat{f} = R^{\rm opt}(\mathbf{y})$ be the solution to (22). We point out (as already mentioned or as a special case of (50)) that $\hat{f} - P_V \hat{f}$ is orthogonal to the space $\ker(L)$. This property completely characterizes

$\hat{f}$ as the element given by (25). Indeed, in view of $\ker(L)^\perp = \mathrm{span}\{u_1, \ldots, u_m\}$, we have

$$\hat{f} - P_V \hat{f} = \sum_{i=1}^m a_i u_i \qquad \text{for some } \mathbf{a} \in \mathbb{R}^m. \tag{28}$$

Taking inner product with $v_1, \ldots, v_n$ leads to $\mathbf{0} = \mathbf{C}^\top \mathbf{a}$. Then, expanding $P_V \hat{f}$ on $(v_1, \ldots, v_n)$, we obtain

$$\hat{f} = \sum_{i=1}^m a_i u_i + \sum_{j=1}^n b_j v_j \qquad \text{for some } \mathbf{b} \in \mathbb{R}^n. \tag{29}$$

Taking inner product with $u_1, \ldots, u_m$ leads to $\mathbf{y} = \mathbf{Ga} + \mathbf{Cb}$ and in turn to $\mathbf{C}^\top \mathbf{G}^{-1} \mathbf{y} = \mathbf{C}^\top \mathbf{G}^{-1} \mathbf{Cb}$ after multiplying by $\mathbf{C}^\top \mathbf{G}^{-1}$. The latter yields the expression for $\mathbf{b}$ given in (26), while the former yields the expression for $\mathbf{a}$ given in (27). □

3 Relation to Supervised Learning

Supervised learning algorithms take data $\mathbf{y} \in \mathbb{R}^m$ as input (while also being aware of the $\mathbf{x}_i$'s) and return functions $\hat{f} \in \mathcal{H}$ as outputs, so they can be viewed as learning/recovery maps $R : \mathbb{R}^m \to \mathcal{H}$. We examine below how some of them compare to the map R^{opt} from Theorem 2.

3.1 Empirical risk minimizations

By design, the outputs $\hat{f}$ returned by these algorithms belong to a hypothesis space chosen in advance from the belief that it provides good approximants for real-life functions. Since this implicit belief parallels the explicit assumption expressed by the model set (5), our Optimal Recovery algorithm and empirical risk minimization algorithms are directly comparable, in that they both depend on a common approximation space/hypothesis class V. With a loss function chosen as a pth power of an ℓ_p-norm for $p \in [1, \infty]$, empirical risk minimization algorithms consist in solving the convex optimization program

$$\underset{f \in \mathcal{H}}{\mathrm{minimize}} \, \|\mathbf{y} - L(f)\|_p^p = \sum_{i=1}^m |y_i - \ell_i(f)|^p \quad \text{s.to } f \in V. \tag{30}$$

In the case $p = 2$ of the square loss, the solution actually reads

$$R^{\mathrm{erm}_2}(\mathbf{y}) = \sum_{j=1}^n \Big((\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top \mathbf{y} \Big)_j v_j, \tag{31}$$

where the matrix $\mathbf{C} \in \mathbb{R}^{m \times n}$ still represents the cross-Gramian introduced in (24).

3.2 Kernel regressions

Kernel regression algorithms usually operate in the setting of Reproducing Kernel Hilbert Spaces (see next section), but they can be phrased for arbitrary Hilbert spaces, too. For instance, the traditional kernel ridge regression consists in solving the following convex optimization problem

$$\underset{f\in\mathcal{H}}{\text{minimize}} \sum_{i=1}^{m}(y_i-\ell_i(f))^2+\gamma\|f\|_{\mathcal{H}}^2 \tag{32}$$

for some parameter $\gamma > 0$. In the limit $\gamma \to 0$, one obtains kernel ridgeless regression, which consists in solving the convex optimization problem

$$\underset{f\in\mathcal{H}}{\text{minimize}}\, \|f\|_{\mathcal{H}} \qquad \text{s.to } \ell_i(f)=y_i, \quad i\in[1:m]. \tag{33}$$

This algorithm fits the training data perfectly and is also known to generalize well in the presence of noise [17].

The crucial observation we wish to bring forward here is that kernel ridgeless regression, although not designed with this intention, is also an Optimal Recovery method. Indeed, (33) appears as the special case of the convex optimization program (22) with the choice $V=\{0\}$. Using Theorem 2, we can retrieve in particular that kernel ridgeless regression is explicitly given by

$$R^{\text{ridgeless}}(\mathbf{y})=\sum_{i=1}^{m}\left(\mathbf{G}^{-1}\mathbf{y}\right)_i u_i. \tag{34}$$

Incidentally, the latter can also be interpreted as the special case where $V=\text{span}\{u_1,\ldots,u_m\}$, since $\hat{f}=R^{\text{ridgeless}}(\mathbf{y})$ is a linear combination of the Riesz representers $u_1,\ldots,u_m$ that satisfy the observation constraint $L(\hat{f})=\mathbf{y}$. In fact, there are more choices for V that leads to kernel ridgeless regression, as revealed below.

Theorem 3. *If the space is $V=\text{span}\{u_i, i\in I\}$ for some subset I of $[1:m]$, then the locally optimal recovery map (22) reduces to kernel ridgeless regression independently of I.*

Proof. Let $V=\text{span}\{u_i, i\in I\}$ for some $I\subseteq[1:m]$ and let $\hat{f}$ be the output of kernel ridgeless regression. According to the proof of Theorem 2, to prove that $\hat{f}$ is the solution to (22), we have to verify that $\hat{f}-P_V\hat{f}\in\ker(L)^\perp$. Since we already know that $\hat{f}=\hat{f}-P_{\{0\}}\hat{f}\in\ker(L)^\perp$ (recall that kernel ridgeless regression is (22) with $\{0\}$ in place of V), it remains to check that $P_V\hat{f}\in\ker(L)^\perp$. This simply follows from $P_V\hat{f}\in\text{span}\{u_i, i\in I\}\subseteq\text{span}\{u_1,\ldots,u_m\}=\ker(L)^\perp$. □

Remark As revealed in [21], the minimization over R of the worst-case relative error

$$\sup_{f\in\mathcal{H}}\frac{\|f-R(L(f))\|_{\mathcal{H}}}{\|f\|_{\mathcal{H}}} \tag{35}$$

coincides with kernel ridge regression. This minimax problem can be generalized to Banach spaces and interpreted as an adversarial zero-sum game. Thus, with the addition of our observation, kernel regressions appear to be optimal in a game-theoretic sense, in a Statistical Learning (average-case) sense, and in an Optimal Recovery (worst-case) sense.

3.3 Spline models

From an Optimal Recovery point-of-view, the success of (33) can be surprising because it seems to use only data and no model assumption. In fact, the model assumption occurs in the objective function being minimized. Procedure (33) favors data-consistent functions which are themselves small. If one preferred to favor data-consistent functions which have small derivatives, one would instead consider, say, the program

$$\underset{f \in W_2^k[0,1]}{\text{minimize}} \, \|f^{(k)}\|_{L_2} \qquad \text{s.to } f(x_i) = y_i, \quad i \in [1:m], \tag{36}$$

with optimization variable f in the Sobolev space $W_2^k[0,1]$. As it turns out, this procedure coincides with the Optimal Recovery method that minimizes the worst-case error over the model set given by $\mathcal{K} = \{f \in W_2^k[0,1] : \|f^{(k)}\|_{L_2} \leq 1\}$ and its solution is known explicitly [5]. With $k = 2$ (where one tries to minimize the strain energy of a curve constrained to pass through a prescribed set of points), the solution is a cubic spline, see [28] for details. For multivariate functions, the solutions to problems akin to (36) are also known explicitly: they are thin plate splines [10]. More generally, minimum-(semi)norm interpolation problems are what define the concept of abstract splines [7].

Remark. When observation error is present, exact interpolation conditions should not be enforced, so it is natural to substitute (22) by a regularized problem similar to (32) but with $\|f - P_V f\|_{\mathcal{H}}^2$ acting as a reguralizer instead of $\|f\|_{\mathcal{H}}^2$. This has already been proposed in [16] under the name Generalized Regularized Least-Squares, of course with a different motivation than Optimal Recovery. In fact, a more general regularized problem where $\|f - P_V f\|_{\mathcal{H}}^2$ gives way to a squared seminorm also appeared in [9]. Such inverse-problem inspired techniques have been applied to statistical learning e.g. in [8], which studied the consistency properties of the associated estimators.

4 Optimal Recovery in Reproducing Kernel Hilbert Spaces

We consider in this section the case where $\mathcal{F} = \mathcal{H}$ is a Hilbert space of functions defined on a domain $\Omega \subseteq \mathbb{R}^d$ for which point evaluations are continuous linear functionals. In other words, we consider a reproducing kernel Hilbert space $\mathcal{H}_K$, where $K : \Omega \times \Omega \to \mathbb{R}$ denotes the kernel characterized, for any $\mathbf{x} \in \Omega$, by

$$f(\mathbf{x}) = \langle K(\mathbf{x}, \cdot), f \rangle \qquad \text{for all } f \in \mathcal{H}_K. \tag{37}$$

In this way, the Riesz representers of points evaluations at $\mathbf{x}_i$'s take the form $u_i = K(\mathbf{x}_i, \cdot)$. Thus, the Gramian of (23) has entries

$$\mathbf{G}_{i,i'} = \langle K(\mathbf{x}_i, \cdot), K(\mathbf{x}_{i'}, \cdot)\rangle = K(\mathbf{x}_{i'}, \mathbf{x}_i), \; i, i' \in [1:m]. \tag{38}$$

As for the cross-Gramian of (24), it has entries

$$\mathbf{C}_{i,j} = v_j(\mathbf{x}_i), \qquad i \in [1:m], \; j \in [1:n], \tag{39}$$

where $(v_1, \ldots, v_n)$ represents a basis for the space V. Some possible choices of K and V are discussed below.

4.1 Choosing the kernel

A kernel that is widely used in many learning problems is the Gaussian kernel given, for some parameter $\sigma > 0$, by

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right), \qquad \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d. \tag{40}$$

The associated infinite-dimensional Hilbert space, explicitly characterized in [20], has orthonormal basis $\{\phi_\alpha, \alpha \in \mathbb{N}_0^d\}$, where

$$\phi_\alpha(\mathbf{x}) = \sqrt{\frac{(1/\sigma^2)^{\alpha_1 + \cdots + \alpha_d}}{\alpha_1! \cdots \alpha_d!}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2\sigma^2}\right) x_1^{\alpha_1} \cdots x_d^{\alpha_d}. \tag{41}$$

4.2 Choosing the approximation space

Since a learning/recovery procedure uses both data and model (maybe implicitly), its performance depends on the interaction between the two. In Optimal Recovery, and subsequently in Information-Based Complexity [26], it is often assumed that the model is fixed and that the user has the ability to choose evaluation points in a favorable way. From another angle, one can view the evaluation points as being fixed but the model could be chosen accordingly. For the applicability of Theorem 2, it is perfectly fine to select an approximation space V depending on $\mathbf{x}_1, \ldots, \mathbf{x}_m$, so long as it does not depend on $y_1, \ldots, y_m$. Thus, one possible choice for the approximation space consists of $V = \mathrm{span}\{K(\mathbf{x}_i, \cdot), i \in I\}$ for some subset $I \subseteq [1:m]$. However, we have seen in Theorem 3 that such a choice invariably leads to kernel ridgeless regression. Another choice for the approximation space is inspired by linear regression, which uses the space $\mathrm{span}\{1, x_1, \ldots, x_d\}$. We do not consider this space verbatim, because its elements (or any polynomial function, for that matter, see [20]) do not belong to the reproducing kernel Hilbert space with Gaussian kernel. Instead, we modify it slightly by multiplying with a decreasing exponential and by allowing for degrees k higher than one, so as to consider the space

$$V = \operatorname{span}\{\phi_\alpha, \alpha_1 + \cdots + \alpha_d \le k\}, \tag{42}$$

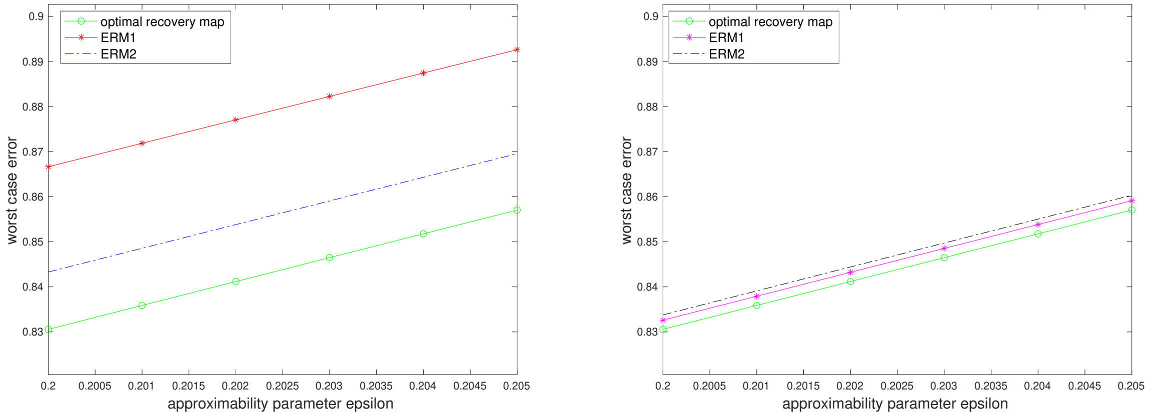
which has dimension $n = \binom{d+k}{d}$. We ignore the coefficients of ϕ_α in numerical experiments, which has no effects on the test error. These ϕ_α's are the so-called 'Taylor features' used in approximation of the Gaussian kernel [4].

5 Experimental Validation

5.1 Comparison of worst-case errors

We first compare worst-case errors for the optimal recovery map (OR) described in Theorem 2 and for empirical risk minimizations defined in (30). They are only considered with $p = 1$ (ERM1) and $p = 2$ (ERM2). The algorithms OR, ERM1, and ERM2 all operate with a specific space V (as a hypothesis class), so direct comparisons can be made by selecting the same V for all these algorithms. According to Theorem 1, when $\mathcal{H}$ is a finite-dimensional Hilbert space, the computation of their worst-case errors is performed by semidefinite programming. Here, we restrict ourselves to the case where V is a randomly generated n-dimensional subspace of $\mathcal{H} = \ell_2^N$, with $n = 20$ and $N = 200$. The observation operator L is also randomly generated by taking the $m = 50$ Riesz representers $u_i \in \mathcal{H}$ as vectors whose entries are independent and uniformly distributed on $[0, 1]$. The observation vector $y \in \mathbb{R}^m$ is obtained by applying L to a randomly generated $f_0 \in \mathcal{H}$ satisfying $\|f_0 - P_V f_0\|_{\mathcal{H}} \le \epsilon_0 := 0.2$. Figure 1(a), which corresponds to a particular realization of V, L, and f_0 and to an approximation parameter ϵ varying from ϵ_0 to $\epsilon_1 := 0.205$, confirms that OR yields the smallest worst-case errors. It also hints at a quasi-linear dependence of the worst-case errors on ϵ and suggests that ERM2 yields smaller worst-case errors than ERM1.

In contrast, keeping the same realization of V, L, and f_0 as above, Figure 1(b) suggests that ERM1 yields smaller worst-case errors than ERM2 when the standard empirical risk minimization (30) is enhanced by replacing the overdemanding constraint $f \in V$ by the constraint $f \in \mathcal{K}$, i.e., $\|f - P_V f\|_{\mathcal{H}} \le \epsilon$. Although the performances are now very close for all algorithms, it has to be noted that in this case running ERM1 and ERM2 requires an a priori knowledge of ϵ while running OR does not.



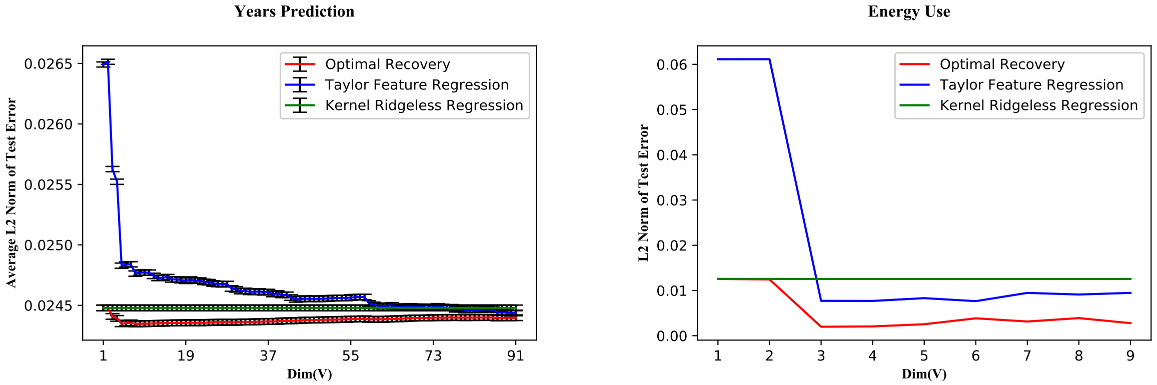
(a) ERM1 and ERM2 with constraint $f \in V$.

(b) ERM1 and ERM2 with constraint $f \in \mathcal{K}$.

Figure 1: Optimal Recovery and Empirical Risk Minimization maps with $p = 1$ and $p = 2$.

5.2 Test errors for non-IID data

In this subsection, we implement the optimal recovery map on two real-world regression datasets, namely Years Prediction and Energy Use. Both of these open-access datasets are widely used for algorithm evaluation and are available on UCI Machine Learning Repository. We focus on the reproducing kernel Hilbert space $\mathcal{H}_K$ associated with Gaussian kernel throughout this experiment. The space V is spanned by a subset of Taylor features of order $k = 1$, see (42), so that $\dim(V)$ goes up to $d+1$, where d is the number of features in the datasets. The basis are generated as described in (23) and (24). To be specific, instances in C are the Taylor features described in (41) evaluated on the observations and instances in G are the kernel function evaluated on the observation pairs. To choose the optimal kernel width, we conduct a grid search. Furthermore, to make the data non-IID, we sort both datasets according to their 5-th feature in a descending order and then select the top 70% as the training set and the bottom 30% as the test set. Note that one can sort by any feature to create the same non-IID condition. Recall by Theorem 2 that the optimal recovery map depends on the Hilbert space $\mathcal{H}_K$ and the subspace V. Therefore, it is natural to compare it to kernel ridgeless regression (34) (in $\mathcal{H}_K$) and Taylor features regression (31) (in V). The test error comparison is presented in Figure 2. Due to the size of Years Prediction dataset, we do not perform kernel ridgeless regression on the full dataset, so we randomly subsample a 5000 subset of the data and repeat the experiment for 40 Monte Carlo simulations to average out the randomness. Therefore, error bars are presented in Figure 2(a) to show the statistical significance. We observe that the optimal recovery map shows promising performance on both datasets. On Years Prediction dataset, Optimal Recovery outperforms kernel ridgeless regression for all $\dim(V)$. On Energy Use dataset, it outperforms kernel ridgeless regression after $\dim(V) = 2$. Also, Taylor features regression in the space V is consistently inferior to the optimal recovery map. The U-shape Optimal Recovery curve in Figure 2(a) demonstrates the trade-off between the compatibility indicator μ and the approximability parameter ϵ.



(a) Test error comparison on Years Prediction

(b) Test error comparison on Energy Use

Figure 2: Optimal Recovery and two benchmark regression algorithms on two benchmark datasets.

6 Conclusion

Generalization guarantees in Statistical Learning are based on the postulate of IID data, the pertinence of which is not guaranteed in all learning environments. In this work, we considered the regression problem (with non-random data) in Hilbert spaces from an Optimal Recovery point-of-view, where the learner aims at minimizing the worst-case error. We first formulated a semidefinite program for calculating the worst-case error of any recovery map in finite-dimensional Hilbert spaces. Then, we provided a closed-form expression for optimal recovery map in the case where the hypothesis class V is a linear subspace of any Hilbert space. The formula coincides with kernel ridgeless regression when $V = \{0\}$ in a reproducing kernel Hilbert space. Our numerical experiments showed that, when $\dim(V) > 0$, Optimal Recovery has the potential to outperform kernel ridgeless regression in the test mean squared error. Our main focus was to provide an algorithmic perspective to Optimal Recovery, whose theory was initiated in the 70's-80's. Our findings revealed interesting connections with current Machine Learning methods. There are many directions to consider in the future, including:

(i) learning the hypothesis space V from the data (instead of incorporating domain knowledge);

(ii) developing Optimal Recovery with noise/error in the observations;[2]

[2] Some results can now be found in [13], in particular, the article [13] uncovers a principled way to choose the parameter of a Tikhonov-like regression.

(iii) studying the overparametrized regime $\dim(V) > m$;

(iv) investigating the case where the hypothesis class V is not a linear space.

Acknowledgment

C. L. and Y. W. are supported by the Texas A&M Triads for Transformation (T3) Program. S. F. was/is partially supported by grants from the NSF (DMS-1622134, DMS-1664803, DMS-2053172) and from the ONR (N00014-20-1-2787). S. F and S. S. also acknowledge NSF grant CCF-1934904.

Appendix

Below, we generalize the result of [3] in two directions. For the first direction, instead of assuming that the target function f_0 itself is well approximated by elements of a set V, we assume that it is some linear transform T applied to f_0 that is well approximated. This translates in the modification (44) of the approximability set. The novelty occurs not for invertible transforms, but for noninvertible ones (e.g. when T represents a derivative, as in (36)). For the second direction, instead of attempting to recover f_0 in full, we assume that we only need to estimate a quantity $Q(f_0)$ depending on f_0, such as its integral. Although we focus on the extreme situations where Q is the identity or where Q is a linear functional, the case of an arbitrary linear map Q is covered. Leaving the introduction of the transform T aside, one useful consequence of the result below is that knowledge of (a basis for) the space V is not needed, since only the values of the $\ell_i(v_j)$'s and $Q(v_j)$'s are required to form $(Q \circ R^{\rm opt})(\mathbf{y})$.

Theorem 4. Let $\mathcal{F}, \mathcal{H}, \mathcal{Z}$ be three normed spaces, $\mathcal{H}$ being a Hilbert space, and let V be a subspace of $\mathcal{H}$. Consider a linear quantity of interest $Q : \mathcal{F} \to \mathcal{Z}$ and a linear map $T : \mathcal{F} \to \mathcal{H}$. For $\mathbf{y} \in \mathbb{R}^m$, define $R^{\rm opt}(\mathbf{y}) \in \mathcal{F}$ as a solution to

$$\underset{f \in \mathcal{F}}{\rm minimize} \, \|Tf - P_V(Tf)\|_{\mathcal{H}} \qquad \text{s.to } L(f) = \mathbf{y}. \tag{43}$$

Then the learning/recovery map $Q \circ R^{\rm opt} : \mathbb{R}^m \to \mathcal{Z}$ is locally optimal over the model set

$$\mathcal{K} = \{f \in \mathcal{F} : {\rm dist}(Tf, V) \le \epsilon\} \tag{44}$$

in the sense that, for any $z \in \mathcal{Z}$,

$$\sup_{f \in \mathcal{K}, L(f) = \mathbf{y}} \|Q(f) - Q \circ R^{\rm opt}(f)\|_{\mathcal{Z}} \le \sup_{f \in \mathcal{K}, L(f) = \mathbf{y}} \|Q(f) - z\|_{\mathcal{Z}}. \tag{45}$$

Proof. Let us introduce the compatibility indicator

$$\mu := \sup_{u \in \ker(L) \setminus \{0\}} \frac{\|Q(u)\|_{\mathcal{Z}}}{\operatorname{dist}(Tu, V)}. \tag{46}$$

Given $\mathbf{y} \in \mathbb{R}^m$, let $\hat{f} = R^{\mathrm{opt}}(\mathbf{y})$ denote the solution to (43). We shall establish (45) by showing on the one hand that

$$\sup_{\substack{f \in \mathcal{K} \\ L(f) = \mathbf{y}}} \|Q(f) - Q(\hat{f})\|_{\mathcal{Z}} \le \mu \big[\epsilon^2 - \|T\hat{f} - P_V(T\hat{f})\|_{\mathcal{H}}^2\big]^{1/2} \tag{47}$$

and on the other hand that, for any $z \in \mathcal{Z}$.

$$\sup_{\substack{f \in \mathcal{K} \\ L(f) = \mathbf{y}}} \|Q(f) - z\|_{\mathcal{Z}} \ge \mu \big[\epsilon^2 - \|T\hat{f} - P_V(T\hat{f})\|_{\mathcal{H}}^2\big]^{1/2}. \tag{48}$$

Let us start with (47). Considering an arbitrary $u \in \ker(L)$, notice that the quadratic expression $t \in \mathbb{R}$ given by

$$\|T(\hat{f} + tu) - P_V(T(\hat{f} + tu))\|_{\mathcal{H}}^2 = \|T\hat{f} - P_V(T\hat{f})\|_{\mathcal{H}}^2 + 2t\langle T\hat{f} - P_V(T\hat{f}), Tu - P_V(Tu)\rangle + \mathcal{O}(t^2) \tag{49}$$

is miminized at the point $t = 0$. This forces the linear term $\langle T\hat{f} - P_V(T\hat{f}), Tu - P_V(Tu)\rangle$ to vanish, in other words

$$\langle T\hat{f} - P_V(T\hat{f}), Tu\rangle = 0 \qquad \text{for any } u \in \ker(L). \tag{50}$$

Now, considering $f \in \mathcal{K}$ such that $L(f) = \mathbf{y}$ written as $f = \hat{f} + u$ for some $u \in \ker(L)$, the fact that $f \in \mathcal{K}$ reads

$$\epsilon^2 \ge \|T(\hat{f} + u) - P_V(T(\hat{f} + u))\|_{\mathcal{H}}^2 = \|T\hat{f} - P_V(T\hat{f})\|_{\mathcal{H}}^2 + \|Tu - P_V(Tu)\|_{\mathcal{H}}^2. \tag{51}$$

Rearranging the latter gives

$$\operatorname{dist}(Tu, V) \le \big[\epsilon^2 - \|T\hat{f} - P_V(T\hat{f})\|_{\mathcal{H}}^2\big]^{1/2}. \tag{52}$$

It remains to take the definition (46) into account in order to bound $\|Q(f) - Q(\hat{f})\|_{\mathcal{Z}} = \|Q(u)\|_{\mathcal{Z}}$ and arrive at (47).

Turning to (48), we consider $u \in \ker(L)$ such that

$$\|Q(u)\|_{\mathcal{Z}} = \mu \operatorname{dist}(Tu, V), \tag{53}$$

$$\|Tu - P_V(Tu)\|_{\mathcal{H}} = \big[\epsilon^2 - \|T\hat{f} - P_V(T\hat{f})\|_{\mathcal{H}}^2\big]^{1/2}. \tag{54}$$

It is clear that $f^{\pm} := \hat{f} \pm u$ both satisfy $L(f^{\pm}) = \mathbf{y}$, while $f^{\pm} \in \mathcal{K}$ follows from

$$\begin{aligned} \|Tf^{\pm} - P_V(Tf^{\pm})\|_{\mathcal{H}}^2 &= \|(T\hat{f} - P_V(T\hat{f})) \pm (Tu - P_V(Tu))\|_{\mathcal{H}}^2 \\ &= \|T\hat{f} - P_V(T\hat{f})\|^2 + \|Tu - P_V(Tu)\|_{\mathcal{H}}^2 = \epsilon^2. \end{aligned} \tag{55}$$

Therefore, for any $z \in \mathcal{Z}$,

$$\begin{aligned}\sup_{\substack{f \in \mathcal{K} \\ L(f)=\mathbf{y}}} \|Q(f) - z\|_{\mathcal{Z}} &\geq \max\{\|Q(f^+) - z\|_{\mathcal{Z}}, \|Q(f^-) - z\|_{\mathcal{Z}}\} \\ &\geq \frac{1}{2}\big(\|Q(f^+) - z\|_{\mathcal{Z}} + \|Q(f^-) - z\|_{\mathcal{Z}}\}\big) \\ &\geq \frac{1}{2}\|Q(f^+ - f^-)\|_{\mathcal{Z}} = \|Q(u)\|_{\mathcal{Z}}. \end{aligned} \tag{56}$$

Taking (53) and (54) into account finishes to prove (48). □

References

[1] A. Beck and Y. C. Eldar. Regularization in regression with bounded noise: A Chebyshev center approach. *SIAM Journal on Matrix Analysis and Applications*, 29(2):606–625, 2007.

[2] M. Belkin, S. Ma, and S. Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549, 2018.

[3] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Data assimilation in reduced modeling. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1–29, 2017.

[4] A. Cotter, J. Keshet, and N. Srebro. Explicit approximations of the Gaussian kernel. *arXiv preprint arXiv:1109.4603*, 2011.

[5] C. de Boor. Best approximation properties of spline functions of odd degree. *Journal of Mathematics and Mechanics*, pages 747–749, 1963.

[6] C. de Boor. Computational aspects of optimal recovery. In *Optimal Estimation in Approximation Theory*, pages 69–91. Springer, 1977.

[7] C. de Boor. Convergence of abstract splines. *Journal of Approximation Theory*, 31(1):80–89, 1981.

[8] E. De Vito, L. Rosasco, A. Caponnetto, U. Giovannini, and F. Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904, 2005.

[9] E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *J. Mach. Learn. Res.*, 5:1363–1390, Dec. 2004.

[10] J. Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer, 1977.

[11] M. Ettehad and S. Foucart. Instances of computational optimal recovery: dealing with observation errors. *arXiv preprint arXiv:2004.00192*, 2020.

[12] S. Foucart. Instances of computational optimal recovery: refined approximability models. *Journal of Complexity*, 62:101503, 2021.

[13] S. Foucart and C. Liao. Optimal recovery from inaccurate data in Hilbert spaces: Regularize, but what of the parameter? *arXiv preprint arXiv:2111.02601*, 2021.

[14] Gurobi Optimization, LLC. Gurobi optimizer reference manual, 2020.

[15] E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[16] W. Li, K.-H. Lee, and K.-S. Leung. Generalized regularized least-squares learning with predefined features in a Hilbert space. In *Advances in Neural Information Processing Systems*, pages 881–888, 2007.

[17] T. Liang and A. Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 2019.

[18] C. A. Micchelli and T. J. Rivlin. A survey of optimal recovery. In *Optimal Estimation in Approximation Theory*, pages 1–54. Springer, 1977.

[19] C. A. Micchelli and T. J. Rivlin. Lectures on optimal recovery. In *Numerical Analysis Lancaster 1984*, pages 21–93. Springer, 1985.

[20] H. Minh. Some properties of Gaussian reproducing kernel Hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32:307–338, 2010.

[21] H. Owhadi, C. Scovel, and F. Schäfer. Statistical numerical approximation. *Notices of the American Mathematical Society*, 66:1608–1617, 2019.

[22] E. W. Packel. Do linear problems have linear optimal algorithms? *SIAM Review*, 30(3):388–403, 1988.

[23] L. Plaskota. *Noisy information and computational complexity*, volume 95. Cambridge University Press, 1996.

[24] I. Pólik and T. Terlaky. A survey of the S-lemma. *SIAM Review*, 49(3):371–418, 2007.

[25] A. Rakhlin and X. Zhai. Consistency of interpolation with Laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pages 2595–2623, 2019.

[26] J. F. Traub. *Information-Based Complexity.* John Wiley and Sons Ltd., 2003.

[27] V. N. Vapnik. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

[28] G. Wahba. *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, 1990.

[29] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.