Learning functions using data-dependent regularization: Representer theorem revisited

Qing Zou[0000−0002−2729−2491]

Applied Mathematics and Computational Sciences
University of Iowa, Iowa City, IA 52242, USA
zou-qing@uiowa.edu

Abstract. We introduce a data-dependent regularization problem which uses the geometry structure of the data to learn functions from incomplete data. We show another proof of the standard representer theorem when introducing the problem. At the end of the paper, two applications in image processing are used to illustrate the function learning framework.

Keywords: function learning · manifold structure · representer theorem.

1 Introduction

1.1 Background

Many machine learning problems involve the learning of multidimensional functions from incomplete training data. For example, the classification problem can be viewed as learning a function whose function values give the classes that the inputs belong to. The direct representation of the function in high-dimensional spaces often suffers from the issue of dimensionality. The large number of parameters in the function representation would translate to the need of extensive training data, which is expensive to obtain. However, researchers found that many natural datasets have extensive structure presented in them, which is usually known as manifold structure. The intrinsic structure of the data can then be used to improve the learning results. Nowadays, assuming data lying on or close to a manifold becomes more and more common in machine learning. It is called manifold assumption in machine learning. Though researchers are not clear about the theoretical reason why the datasets have manifold structure, it is useful for supervised learning and it gives excellent performance. In this work, we will exploit the manifold structure to learn functions from incomplete training data.

1.2 A Motivated Example

One of the main problems in numerical analysis is function approximation. During the last several decades, researchers usually considered the following problem

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_23

to apply the theory of function approximation to real-world problems:

$$\min_f ||Lf||^2, \quad s.t. \quad f(x_i) = y_i, \tag{1}$$

where L is some linear operator, $\{(x_i, y_i)\}_{i=1}^n \subset X \times \mathbb{R}$ are n accessible observations and $X \subset \mathbb{R}^d, d \geq 1$ is the input space. We can use the method of Lagrange multiplier to solve Problem (1). Assume that the searching space for the function f is large enough (for example $\mathcal{L}_2$ space). Then the Lagrangian function $C(f)$ is given by

$$C(f) = \langle Lf, Lf \rangle + \sum_{i=1}^n \lambda_i (f(x_i) - y_i).$$

Taking the gradient of the Lagrangian function w.r.t. the function f gives us

$$\begin{aligned} C'(f) &= \lim_{\varepsilon \to 0} \frac{C(f + \varepsilon f) - C(f)}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{2\varepsilon \langle Lf, Lf \rangle + \varepsilon \sum \lambda_i \varepsilon f(x_i)}{\varepsilon} \\ &= 2\langle Lf, Lf \rangle + \sum_{i=1}^n \lambda_i f(x_i). \end{aligned}$$

Setting $C'(f) = 0$, we have

$$2\langle Lf, Lf \rangle = -\sum_{i=1}^n \lambda_i f(x_i) = -\sum_{i=1}^n \lambda_i \langle f(x), \delta(x - x_i) \rangle,$$

where $\delta(\cdot - x)$ is the delta function. Suppose L^* is the adjoint operator of L. Then we have

$$2\langle f(x), (L + L^*)f \rangle = \langle f(x), -\sum_{i=1}^n \lambda_i \delta(x - x_i) \rangle,$$

which gives us $2(L+L^*)f = -\sum_{i=1}^n \lambda_i \delta(x-x_i)$. This implies $f = \sum_{i=1}^n a_i \ell(x, x_i)$, for some a_i and $\ell(\cdot, \cdot)$.

1.3 Kernels and Representer Theorem

As machine learning develops fast these years, kernel methods [1] have received much attentions. Researchers found that working in the original data space is somehow not well-performed. So, we would like to map the data to a high dimensional space (feature space) using some non-linear mapping (feature map). Then we can do a better job (e.g. classification) in the feature space. When we talk about feature map, one concept that is unavoidable to mention is the kernel, which easily speaking is the inner product of the features. With a kernel (positive definite), we can then have a corresponding reproducing kernel Hilbert space (RKHS) [2] $\mathcal{H}_K$. We can now solve the problem that is similar to (1) in the RKHS:

$$\min_{f \in \mathcal{H}_K} ||f||^2_{\mathcal{H}_K} \quad s.t. \quad f(x_i) = y_i.$$

A more feasible way is to consider a regularization problem in the RKHS:

$$\min_{f\in\mathcal{H}_K} ||f(x_i)-y_i||^2+\lambda||f||^2_{\mathcal{H}_K}. \tag{2}$$

Then the searching space of f becomes $\mathcal{H}_K$, which is a Hilbert space. Before solving Problem (2), we would like to recall some basic concepts about the RKHS. Suppose we have a positive definite kernel $K: X\times X\to\mathbb{R}$, i.e.,

$$\sum_{i=1}^n\sum_{j=1}^n a_ia_jK(x_i,x_j)\geq 0,\quad n\in\mathbb{N},\;\; x_1,\cdots,x_n\in X,\;\; a_1,\cdots,a_n\in\mathbb{R},$$

then $\mathcal{H}_K$ is the Hilbert space corresponding to the kernel $K(\cdot,\cdot)$. It is defined by all the possible linear combination of the kernel $K(\cdot,\cdot)$, i.e., $\mathcal{H}_K=\text{span}\{K(\cdot,\cdot)\}$. Thus, for any $f(\cdot)\in\mathcal{H}_K$, there exists x_i and α_i such that

$$f(\cdot)=\sum_i\alpha_iK(\cdot,x_i).$$

Since $\mathcal{H}_K$ is a Hilbert space, it is equipped with an inner product. The principle to define the inner product is to let $\mathcal{H}_K$ have representer $K(\cdot,x)$ and the representer performs like the delta function for functions in $\mathcal{L}_2$ (note that delta function is not in $\mathcal{L}_2$). In other word, we want to have a similar result to the following formula:

$$f(x)=\langle f(\cdot),\delta(\cdot-x)\rangle_{\mathcal{L}_2}.$$

This is called reproducing relation or reproducing property. In $\mathcal{H}_K$, we want to define the inner product so that we have the reproducing relation in $\mathcal{H}_K$:

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

To achieve this goal, we can define

$$\begin{aligned}\langle f,g\rangle_{\mathcal{H}_K}&=\langle\sum_i\alpha_iK(\cdot,x_i),\sum_j\beta_jK(\cdot,x_j)\rangle_{\mathcal{H}_K}\\&=:\sum_i\sum_j\alpha_i\beta_jK(x_i,x_j).\end{aligned}$$

Then we have

$$\langle K(\cdot,x),K(\cdot,y)\rangle_{\mathcal{H}_K}=K(x,y)$$

With the kernel, the feature map $\Phi_\cdot(x)$ can be defined as

$$\Phi_\cdot(x)=K(\cdot,x).$$

Having these knowledge about the RKHS, we can now look at the solution of Problem (2). It can be characterized by the famous conclusion named representer theorem, which states that the solution of Problem (2) is

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

DOI: 10.1007/978-3-030-50420-5_23

The standard proof of the representer theorem is well-known and can be found in many literatures, see for example [3,4]. While the drawback of the standard proof is that the proof did not provide the expression of the coefficients α_i. In the first part of this work, we will provide another proof of the representer theorem. As a by-product, we can also build the relation between Problem (1) and Problem (2).

2 Another Proof of Representer Theorem

To give another proof of the representer theorem, we first build some relations between $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ and $\langle \cdot, \cdot \rangle_{\mathcal{L}_2}$. We endow the dataset X with a measure μ. Then the corresponding $\mathcal{L}_2(X)$ inner product is given by

$$\langle f, g \rangle_{\mathcal{L}_2} = \int_X f \cdot g d\mu.$$

Consider an operator L on f with respect to the kernel K:

$$Lf(x) = \int_X f(y) K(x,y) d\mu, \tag{3}$$

which is the Hilbert-Schmidt integral operator [5]. This operator is self-adjoint, bounded and compact. By the spectral theorem [6], we can obtain that the eigenfunctions $e_1(x), e_2(x), \cdots$ of the operator will form an orthonormal basis of $\mathcal{L}_2(X)$, i.e.,

$$\langle e_i, e_j \rangle_{\mathcal{L}_2} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}.$$

With the operator L defined as (3), we can look at the relations between $\langle \cdot, \cdot \rangle_{\mathcal{L}_2(X)}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$. Suppose $e_i(x)$ are the eigenfunctions of the operator L and λ_i are the corresponding eigenvalues, then

$$\langle K(x,y), e_i(y) \rangle_{\mathcal{L}_2(X)} = \int_X e_i(y) K(x,y) d\mu = \lambda_i e_i(x). \tag{4}$$

But by the reproducing relation, we have

$$\langle K(x,y), e_i(y) \rangle_{\mathcal{H}_K} = e_i(x).$$

Now, let us look at how to represent $K(x,y)$ by the eigenfunctions. We have

$$K(x,y) = \sum_i \lambda_i e_i(x) e_i(y),$$

and λ_i can be computed by

$$\lambda_i = \int_X \int_X K(x,y) e_i(x) e_i(y) dx dy.$$

To see $K(x,y) = \sum_i \lambda_i e_i(x) e_i(y)$, we can just plug it into (4) to verify it:

$$\langle K(x,y), e_i(y) \rangle_{\mathcal{L}_2(X)} = \int_X e_i(y) \sum_j \lambda_j e_j(x) e_j(y) d\mu(y)$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_23

$$= \sum_j \lambda_j \int_X e_j(x) e_i(y) e_j(y) d\mu(y) = \sum_j \lambda_j e_j(x) \int_X e_i(y) e_j(y) dy = \lambda_i e_i(x).$$

Since the eigenfunctions of L form an orthogonal basis of $\mathcal{L}_2(X)$, then for any $f \in \mathcal{L}_2$, it can be written as $f = \sum_i a_i e_i(x)$. So we have

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

While for the $\mathcal{L}_2$ norm, we have

$$\langle K(x,\cdot), f(\cdot)\rangle_{\mathcal{L}_2(X)} = \langle K(x,\cdot), \sum_i a_i e_i(\cdot)\rangle_{\mathcal{L}_2(X)}$$
$$= \sum_i a_i \langle K(x,\cdot), e_i(\cdot)\rangle_{\mathcal{L}_2(X)} = \sum_i a_i \lambda_i e_i(x).$$

Next we show that the orthonormal basis $e_i(x)$ are within $\mathcal{H}_K$. Note that

$$e_i(x) = \langle K(x,\cdot), e_i(\cdot)\rangle_{\mathcal{H}_K} = \langle \sum_j \lambda_j e_j(x) e_j(\cdot), e_i(\cdot)\rangle_{\mathcal{H}_K},$$

which implies

$$e_i(x) = \sum_j \lambda_j e_j(x) \langle e_j(\cdot), e_i(\cdot)\rangle_{\mathcal{H}_K}.$$

So we can get

$$\langle e_j, e_i\rangle_{\mathcal{H}_K} = \begin{cases} 0, & i \neq j \\ \frac{1}{\lambda_i}, & i = j \end{cases} < \infty.$$

Therefore, we get $e_i(x) \in \mathcal{H}_K$.

We now need to investigate that for any $f = \sum_i a_i e_i(x) \in \mathcal{L}_2(X)$, when will we have that $f \in \mathcal{H}_K$. To let $f \in \mathcal{H}_K$, we need to have $||f||^2_{\mathcal{H}_K} \leq \infty$. So

$$||f||^2_{\mathcal{H}_K} = \langle f, f\rangle_{\mathcal{H}_K} = \langle \sum_i a_i e_i(x), \sum_i a_i e_i(x)\rangle_{\mathcal{H}_K}$$
$$= \sum_i a_i^2 \langle e_j, e_i\rangle_{\mathcal{H}_K} = \sum_i a_i^2 \cdot \frac{1}{\lambda_i} < \infty.$$

This means that to let $f = \sum_i a_i e_i(x) \in \mathcal{H}_K$, we need to have $\sum_i \frac{a_i^2}{\lambda_i} < \infty$ [7].

Combining all these analysis, we can then get the following relation between $\langle\cdot,\cdot\rangle_{\mathcal{L}_2(X)}$ and $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$:

$$\langle f, g\rangle_{\mathcal{L}_2(X)} = \langle L^{1/2} f, L^{1/2} g\rangle_{\mathcal{H}_K}, \quad \forall f, g \in \mathcal{H}_K, \quad L = L^{1/2} \circ L^{1/2}.$$

According to which, we can have another proof of the representer theorem.

DOI: 10.1007/978-3-030-50420-5_23

Proof. Suppose $e_1, e_2, \cdots$ are eigenfunctions of the operator L. Then we can write the solution as $f^* = \sum_i a_i e_i(x)$. To let $f^* \in \mathcal{H}_K$, we require $\sum_i \frac{a_i^2}{\lambda_i} < \infty$.

We consider here a more general form of Problem (2):

$$\min_{f \in \mathcal{H}_K} \sum_{i=1}^{n} E\left((x_i, y_i), f(x_i)\right) + \lambda ||f||^2_{\mathcal{H}_K},$$

where $E(\cdot, \cdot)$ is the error function which is differentiable with respect to each a_i. We would use the tools in $\mathcal{L}_2(X)$ space to get the solution.

The cost function of the regularization problem is

$$C(f) = \sum_{i=1}^{n} E\left((x_i, y_i), f(x_i)\right) + \lambda ||f||^2_{\mathcal{H}_K}.$$

By substituting f^* into the cost function, we have

$$C(f^*) = \sum_{i=1}^{n} E\left((x_i, y_i), \sum_j a_j e_j(x_i)\right) + \lambda ||f^*||^2_{\mathcal{H}_K}.$$

Since

$$||f^*||^2_{\mathcal{H}_K} = ||\sum_i a_i e_i(x)||^2_{\mathcal{H}_K} = \langle \sum_i a_i e_i(x), \sum_i a_i e_i(x) \rangle_{\mathcal{H}_K} = \sum_i \frac{a_i^2}{\lambda_i} (< \infty),$$

differentiating $C(f^*)$ w.r.t. each a_i and setting it equal to zero gives

$$\frac{\partial C(f^*)}{\partial a_k} = \sum_{i=1}^{n} e_k(x_i) \partial_2 E\left((x_i, y_i), \sum_j a_j e_j(x_i)\right) + 2\lambda \frac{a_k}{\lambda_k} = 0.$$

Solving a_k, we get

$$a_k = -\frac{\lambda_k}{2\lambda} \sum_{i=1}^{n} e_k(x_i) \partial_2 E\left((x_i, y_i), f^*\right).$$

Since $f^* = \sum_k a_k e_k(x)$, we have

$$f^* = \sum_k \left(-\frac{\lambda_k}{2\lambda} \sum_{i=1}^{n} e_k(x_i) \partial_2 E\left((x_i, y_i), f^*\right) \right) e_k(x)$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_23

$$
\begin{aligned}
&= -\frac{1}{2\lambda}\sum_{i=1}^{n}\left(\sum_{k}\lambda_k e_k(x_i)e_k(x)\partial_2 E\left((x_i, y_i), f^*\right)\right) \\
&= -\frac{1}{2\lambda}\sum_{i=1}^{n} K(x, x_i)\cdot \partial_2 E\left((x_i, y_i), f^*\right) \\
&= \sum_{i=1}^{n}\underbrace{\left(-\frac{1}{2\lambda}\partial_2 E\left((x_i, y_i), f^*\right)\right)}_{:=\alpha_i}\cdot K(x, x_i) =: \sum_{i=1}^{n}\alpha_i K(x, x_i).
\end{aligned}
$$

This proves the representer theorem.

Note that this result not only proves the representer theorem, but also gives the expression of the coefficients α_i.

With the operator L, we can also build a relation between Problem (1) and Problem (2). Define the operator in Problem (1) to be the inverse of the Hilbert-Schmidt Integral operator. The discussion on the inverse of the Hilbert-Schmidt Integral operator can be found in [8]. Note that for the delta function, we have

$$
L\delta(x, x_i) = \int_X \delta(y, x_i)K(x, y)dy = K(x, x_i).
$$

Then the solution of Problem (1) becomes $2(2L^{-1})f = -\sum_{i=1}^{n}\lambda_i\delta(x, x_i)$. So we have $L^{-1}f = -\sum_{i=1}^{n}\frac{\lambda_i}{4}\delta(x, x_i)$. Applying L on both sides gives

$$
L(L^{-1}f) = \sum_{i=1}^{n}(-\frac{\lambda_i}{4})L\delta(x, x_i).
$$

By which we obtain

$$
f = \sum_{i=1}^{n}(-\frac{\lambda_i}{4})K(x, x_i) =: \sum_{i=1}^{n}\beta_i K(x, x_i).
$$

3 Data-dependent Regularization

So far, we have introduced the standard representer theorem. While as we discussed at the very beginning, many natural datasets have the manifold structure presented in them. So based on the classical Problem (2), we would like to introduce a new learning problem which exploits the manifold structure of the data. We call it the data-dependent regularization problem. Regularization problem has a long history going back to Tikhonov [9]. He proposed the Tikhonov regularization to solve the ill-posed inverse problem.

To exploit the manifold structure of the data, we can then divide a function into two parts: the function restricted on the manifold and the function restricted outside the manifold. So the problem can be formulated as

$$
\min_{f\in\mathcal{H}_K} ||f(x_i) - y_i||^2 + \alpha||f_1||_{\mathcal{M}}^2 + \beta||f_2||_{\mathcal{M}^c}^2, \tag{5}
$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_23

where $f_1 = f|_{\mathcal{M}}$ and $f_2 = f|_{\mathcal{M}^c}$. The norms $||\cdot||_{\mathcal{M}}$ and $||\cdot||_{\mathcal{M}^c}$ will be explained later in details. α and β are two parameters which control the degree for penalizing the energy of the function on the manifold and outside the manifold. We will show later that by controlling the two balancing parameters (set $\alpha = \beta$), the standard representer theorem is a special case of Problem (5).

We now discuss something about the functions f_1 and f_2. Consider the ambient space $X \subset \mathbb{R}^n$ (or $\mathbb{R}^n$) and a positive definite kernel K. Let us first look at the restriction of K to the manifold $\mathcal{M} \subset X$. The restriction is again a positive definite kernel [2] and it will then have a corresponding Hilbert space. We consider the relation between the RKHS $\mathcal{H}_K$ and the restricted RKHS to explain the norms $||\cdot||_{\mathcal{M}}$ and $||\cdot||_{\mathcal{M}^c}$.

Lemma 1 ([10]). *Suppose $K : X \times X \to \mathbb{R}$ (or $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$) is a positive definite kernel. Let $\mathcal{M}$ be a subset of X (or $\mathbb{R}^n$). $\mathcal{F}(\mathcal{M})$ denote all the functions defined on $\mathcal{M}$. Then the RKHS given by the restricted kernel $K_1 : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is*

$$\mathcal{H}_1(\mathcal{M}) = \{f_1 \in \mathcal{F}(\mathcal{M}) : f_1 = f|_{\mathcal{M}} \text{ for some } f \in \mathcal{H}_K\} \tag{6}$$

with the norm defined as

$$||f_1||_{\mathcal{M}} =: \min\{||f||_{\mathcal{H}_K} : f \in \mathcal{H}_K, f|_{\mathcal{M}} = f_1\}.$$

Proof. Define the set

$$\mathcal{S}(\mathcal{M}) =: \{f_r \in \mathcal{F}(\mathcal{M}) : \exists f \in \mathcal{H}_K \quad s.t. \quad f_r = f|_{\mathcal{M}}\}.$$

We first show that the set $A = \{||f||_{\mathcal{H}_K} : f \in \mathcal{H}_K, f|_{\mathcal{M}} = f_r\}$ has a minuma for any $f_r \in \mathcal{S}(\mathcal{M})$. Choose a sequence $\{f_i\}_{i=1}^{\infty} \subset \mathcal{H}_K$. Then the sequence is bounded because the space $\mathcal{H}_K$ is a Hilbert space. It is reasonable to assume that $\{f_i\}_{i=1}^{\infty}$ is weakly convergent because of the Banach-Alaoglu theorem [11]. By the weakly convergence, we can obtain pointwise convergence according to the reproducing property. So the limit of the sequence $\{f_i\}_{i=1}^{\infty}$ attains the minima.

We further define $||f_r||_{\mathcal{S}(\mathcal{M})} = \min A$. We show that $\big(\mathcal{S}(\mathcal{M}), ||\cdot||_{\mathcal{S}(\mathcal{M})}\big)$ is a Hilbert space by the parallelogram law. In other word, we are going to show that

$$2(||f_1||^2_{\mathcal{S}(\mathcal{M})} + ||g_1||^2_{\mathcal{S}(\mathcal{M})}) = ||f_1 + g_1||^2_{\mathcal{S}(\mathcal{M})} + ||f_1 - g_1||^2_{\mathcal{S}(\mathcal{M})}, \quad \forall f_1, g_1 \in \mathcal{S}(\mathcal{M}).$$

Since we defined $||f_r||_{\mathcal{S}(\mathcal{M})} = \min A$. Then for all $f_1, g_1 \in \mathcal{S}(\mathcal{M})$, there exists $f, g \in \mathcal{H}_K$ such that

$$2(||f_1||^2_{\mathcal{S}(\mathcal{M})} + ||g_1||^2_{\mathcal{S}(\mathcal{M})}) \leq 2(||f||^2_{\mathcal{H}_K} + ||g||^2_{\mathcal{H}_K}) = ||f + g||_{\mathcal{H}_K} + ||f - g||_{\mathcal{H}_K}.$$

By the definition of $\mathcal{S}(\mathcal{M})$, we can choose f_1, g_1such that

$$||f_1 + g_1||^2_{\mathcal{S}(\mathcal{M})} = ||f + g||^2_{\mathcal{H}_K}$$

and

$$||f_1 - g_1||^2_{\mathcal{S}(\mathcal{M})} = ||f - g||^2_{\mathcal{H}_K}.$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_23

Thus, we have

$$2(||f_1||^2_{\mathcal{S}(\mathcal{M})} + ||g_1||^2_{\mathcal{S}(\mathcal{M})}) \leq ||f_1 + g_1||^2_{\mathcal{S}(\mathcal{M})} + ||f_1 - g_1||^2_{\mathcal{S}(\mathcal{M})}.$$

For the reverse inequality, we first choose f_1, g_1 such that $||f_1||^2_{\mathcal{S}(\mathcal{M})} = ||f||^2_{\mathcal{H}_K}$ and $||g_1||^2_{\mathcal{S}(\mathcal{M})} = ||g||^2_{\mathcal{H}_K}$. Then

$$\begin{aligned} &2(||f_1||^2_{\mathcal{S}(\mathcal{M})} + ||g_1||^2_{\mathcal{S}(\mathcal{M})}) = 2(||f||^2_{\mathcal{H}_K} + ||g||^2_{\mathcal{H}_K}) \\ =&||f+g||^2_{\mathcal{H}_K} + ||f-g||^2_{\mathcal{H}_K} \geq ||f_1 + g_1||^2_{\mathcal{S}(\mathcal{M})} + ||f_1 - g_1||^2_{\mathcal{S}(\mathcal{M})}. \end{aligned}$$

Therefore, we get

$$2(||f_1||^2_{\mathcal{S}(\mathcal{M})} + ||g_1||^2_{\mathcal{S}(\mathcal{M})}) = ||f_1 + g_1||^2_{\mathcal{S}(\mathcal{M})} + ||f_1 - g_1||^2_{\mathcal{S}(\mathcal{M})}.$$

Next, we show (6) by showing that for all $f_r \in \mathcal{S}(\mathcal{M})$ and $x \in \mathcal{M}$,

$$f_r(x) = \langle f_r(\cdot), K_1(\cdot, x)\rangle_{\mathcal{S}(\mathcal{M})},$$

where $K_1 = K|_{\mathcal{M}\times\mathcal{M}}$.

Choose $f \in \mathcal{H}_K$ such that $f_r = f|_{\mathcal{M}}$ and $||f_r||_{\mathcal{S}(\mathcal{M})} = ||f||_{\mathcal{H}_K}$. This is possible because of the analysis above. Specially, we have

$$||K_1(\cdot, x)||_{\mathcal{S}(\mathcal{M})} = ||K(\cdot, x)||_{\mathcal{H}_K}, \quad \forall x \in \mathcal{M}.$$

Now, for any function $f \in \mathcal{H}_K$ such that $f|_{\mathcal{M}} = 0$, we have

$$\begin{aligned} &||K(\cdot, x) + f||^2_{\mathcal{H}_K} = ||K(\cdot, x)||^2_{\mathcal{H}_K} + ||f||^2_{\mathcal{H}_K} + 2\langle f, K(\cdot, x)\rangle_{\mathcal{H}_K} \\ =&||K(\cdot, x)||^2_{\mathcal{H}_K} + ||f||^2_{\mathcal{H}_K} + 2f(x) = ||K(\cdot, x)||^2_{\mathcal{H}_K} + ||f||^2_{\mathcal{H}_K}. \end{aligned}$$

Thus,

$$\langle f_r(\cdot), K_1(\cdot, x)\rangle_{\mathcal{S}(\mathcal{M})} = \langle f, K(\cdot, x)\rangle_{\mathcal{H}_K} = f(x) = f_r(x), \quad \forall x \in \mathcal{M}.$$

This completes the proof of the lemma.

With this lemma, the solution of Problem (5) then becomes easy to obtain. By the representer theorem we mentioned before, we know that the function satisfies

$$\min_{f \in \mathcal{H}_K} ||f(x_i) - y_i||^2 + \lambda ||f||^2_{\mathcal{H}_K}$$

is $f = \sum_{i=1}^n a_i K(x, x_i)$. Since we have

$$||f_1||^2_{\mathcal{M}} = \min\{||f||_{\mathcal{H}_K} : f|_{\mathcal{M}} = f_1\},$$

$$||f_2||^2_{\mathcal{M}^c} = \min\{||f||_{\mathcal{H}_K} : f|_{\mathcal{M}^c} = f_2\}.$$

Thus, we can conclude that is solution of (5) is exactly

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

DOI: 10.1007/978-3-030-50420-5_23

where the coefficients a_i are controlled by the parameters α and β.

With the norms $||\cdot||_{\mathcal{M}}$ and $||\cdot||_{\mathcal{M}^c}$ being well-defined, we would like to seek the relation between $||\cdot||_{\mathcal{M}}$, $||\cdot||_{\mathcal{M}^c}$ and $||\cdot||_{\mathcal{H}_K}$. Before stating the relation, we would like to restate some of the notations to make the statement more clear. Let

$$K_1 = K|_{\mathcal{M}\times\mathcal{M}}\,, \qquad K_2 = K|_{(\mathcal{M}\times\mathcal{M})^c}$$

and

$$\mathcal{H}_1(\mathcal{M}) = \{f_1 \in \mathcal{F}(\mathcal{M}) : f_1 = f|_{\mathcal{M}} \text{ for some } f \in \mathcal{H}_K\},$$

$$||f_1||_{\mathcal{M}} =: \min\{||f||_{\mathcal{H}_K} : f \in \mathcal{H}_K,\, f|_{\mathcal{M}} = f_1\}.$$

$$\mathcal{H}_2(\mathcal{M}^c) = \{f_2 \in \mathcal{F}(\mathcal{M}^c) : f_2 = f|_{\mathcal{M}^c} \text{ for some } f \in \mathcal{H}_K\},$$

$$||f_2||_{\mathcal{M}^c} =: \min\{||f||_{\mathcal{H}_K} : f \in \mathcal{H}_K,\, f|_{\mathcal{M}^c} = f_2\}.$$

To find the relation between $||\cdot||_{\mathcal{M}}$, $||\cdot||_{\mathcal{M}^c}$ and $||\cdot||_{\mathcal{H}_K}$, we need to pullback the restricted kernel K_1 and K_2 to the original space. To do so, define

$$K_1^p = \begin{cases} K_1, & (x,y) \in \mathcal{M}\times\mathcal{M} \\ 0, & (x,y) \in (\mathcal{M}\times\mathcal{M})^c \end{cases}.$$

$$K_2^p = \begin{cases} K_2, & (x,y) \in (\mathcal{M}\times\mathcal{M})^c \\ 0, & (x,y) \in \mathcal{M}\times\mathcal{M} \end{cases}.$$

Then we have $K = K_1^p + K_2^p$. The corresponding Hilbert spaces for K_1^p and K_2^p are

$$\mathcal{H}_1^p(\mathcal{M}) = \{f_1^p \in \mathcal{F}(X) :\, f_1^p|_{\mathcal{M}} = f_1,\, f_1^p|_{\mathcal{M}^c} = 0\},$$

$$\mathcal{H}_2^p(\mathcal{M}^c) = \{f_2^p \in \mathcal{F}(X) :\, f_2^p|_{\mathcal{M}^c} = f_2,\, f_2^p|_{\mathcal{M}} = 0\}.$$

It is straightforward to define that

$$||f_1^p||_{\mathcal{H}_{K_1^p}} = ||f_1||_{\mathcal{M}},$$

$$||f_2^p||_{\mathcal{H}_{K_2^p}} = ||f_2||_{\mathcal{M}^c}.$$

The following lemma shows the relation between $||\cdot||_{\mathcal{H}_{K_1^p}}$, $||\cdot||_{\mathcal{H}_{K_2^p}}$ and $||\cdot||_{\mathcal{H}_K}$, which also reveals the relation between $||\cdot||_{\mathcal{M}}$, $||\cdot||_{\mathcal{M}^c}$ and $||\cdot||_{\mathcal{H}_K}$ by Moore-Aronszajn theorem [12].

Lemma 2. *Suppose $K_1, K_2 : Y\times Y \to \mathbb{R}$ (or $\mathbb{R}^n\times\mathbb{R}^n \to \mathbb{R}$) are two positive definie kernels. If $K = K_1 + K_2$, then*

$$\mathcal{H}_K = \{f_1 + f_2 : f_1 \in \mathcal{H}_{K_1}, f_2 \in \mathcal{H}_{K_2}\}$$

is a Hilbert space with the norm defined by

$$||f||^2_{\mathcal{H}_K} = \min_{f_1\in\mathcal{H}_{K_1}, f_2\in\mathcal{H}_{K_2}, f=f_1+f_2} ||f_1||^2_{\mathcal{H}_{K_1}} + ||f_2||^2_{\mathcal{H}_{K_2}}.$$

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_23

The idea of the proof of this lemma is exactly the same as the one for Lemma 1. Thus we omit it here.

A direct corollary of this lemma is:

Corollary 1. *Under the assumption of Lemma 2, if the functions in $\mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ have no functions except for zero function in common. Then the norm of $\mathcal{H}_K$ is given simply by*

$$||f||^2_{\mathcal{H}_K} = ||f_1||^2_{\mathcal{H}_{K_1}} + ||f_2||^2_{\mathcal{H}_{K_2}}.$$

If we go back to our scenario, we can get the following result by Corollary 1:

$$||f||^2_{\mathcal{H}_K} = ||f_1||^2_{\mathcal{M}} + ||f_2||^2_{\mathcal{M}^c}.$$

This means that if we set $\alpha = \beta$ in Problem (5), it will reduce to Problem (2). Therefore, the standard representer theorem is a special case of our data-dependent regularization problem (5).

4 Applications

As we said in the introduction part, many engineering problems can be viewed as learning multidimensional functions from incomplete data. In this section, we would like to show two applications of functions learning: image interpolation and patch-based iamge denoising.

4.1 Image Interpolation

Image interpolation tries to best approximate the color and intensity of a pixel based on the values at surrounding pixels. See Fig. 1 for illustration. From function learning perspective, image interpolation is to learn a function from the known pixels and their corresponding positions.

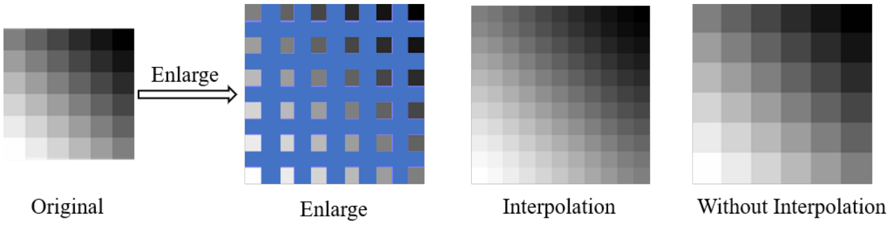


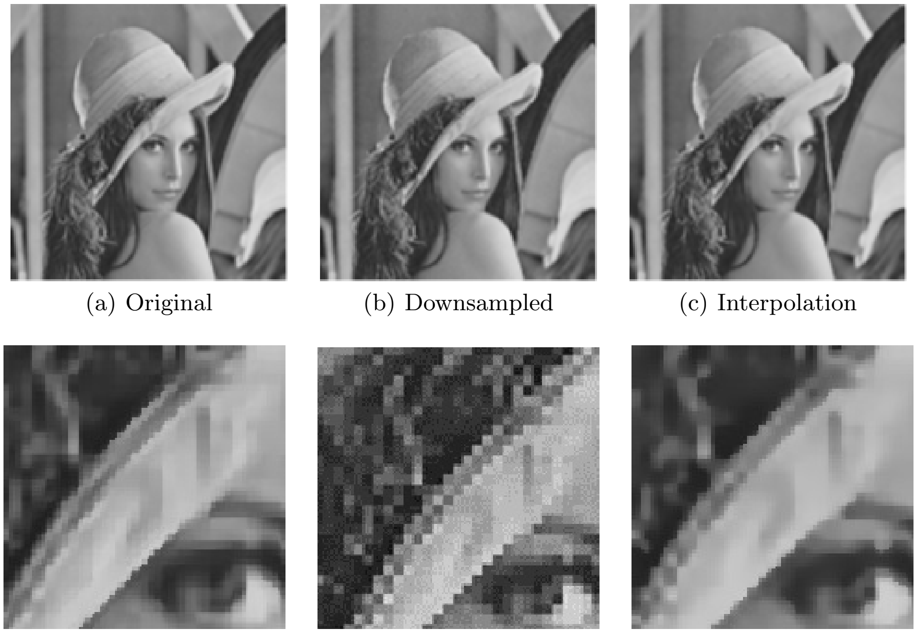
Fig. 1. Illustration of image interpolation. The size of the original image is 6 × 6. We want to enlarge it as an 11 × 11 image. Then the blue shaded positions are unknown. Using image interpolation, we can find the values of these positions.

We would like to use the Lena image as shown in Fig. 2 (a) to give an example of image interpolation utilizing the proposed framework. The zoomed image is shown in Fig. 2 (d). In the image interpolation example, the two balancing parameters are set to be the same and the Laplacian kernel [13] is used:

$$K(x, y) = \exp\left(-\frac{||x - y||}{\sigma}\right).$$

Note that we can also use other kernels, for example, polynomial kernel and Gaussian kernel to proceed image interpolation. Choosing the right kernel is an interesting problem and we do not have enough space to compare different kernels in this paper.

In Fig. 2 (b), we downsampled the original image by a factor of 3 in each direction. The zoomed image is shown in Fig. 2 (e). The interpolation result with the zoomed image are shown in Fig. 2 (c) and Fig. 2 (f).


(a) Original (b) Downsampled (c) Interpolation

(d) Original zoomed (e) Downsampled zoomed (f) Interpolation zoomed

Fig. 2. Illustration of image interpolation. The original image is downsampled by a factor of 3 in each direction. We use the proposed function learning framework to obtain the interpolation function from downsampled image. From the results, we can see that the proposed framework works for image interpolation.

4.2 Patch-based Image Denoising

From the function learning point of view, the patch-based image denoising problem can be viewed as learning a function from noisy patches to their “noise-free” centered pixels. See Fig. 3 for illustration.

In the patch-based image denoising application, we use the Laplacian kernel as well. We assume that the noisy patches are lying close to some manifold so we set the balancing parameter which controls the energy outside the manifold to be large enough. We use the images in Fig. 4 as known data to learn the function. Then for a given noisy image, we can use the learned function to do image denoising. To speed up the learning process, we randomly choose only 10% of the known data to learn the function.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50420-5_23

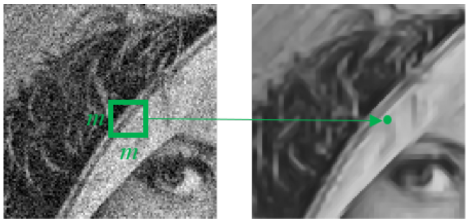
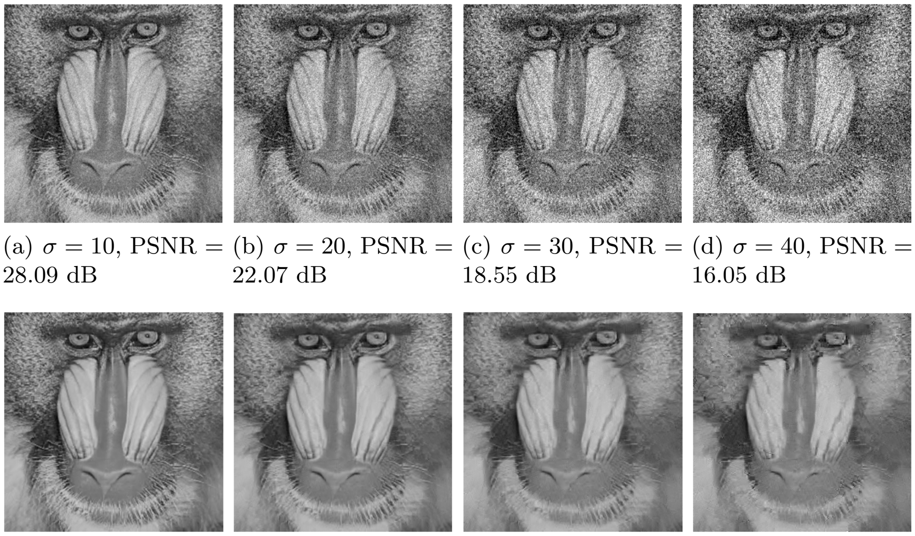


Fig. 3. Illustration of patch-based image denoising. It can be viewed as learning a function from the $m \times m$ noisy patches to the centered clean pixels.

(a) Barbara (b) Couple (c) House (d) hill

Fig. 4. Four training images. We use noisy images and clean pixels to learn the denoising function.

We use the image Baboon to test the learned denoising function. The denoising results are shown in Fig. 5. Each column shows the result corresponding to one noise level.



(a) $\sigma = 10$, PSNR = 28.09 dB (b) $\sigma = 20$, PSNR = 22.07 dB (c) $\sigma = 30$, PSNR = 18.55 dB (d) $\sigma = 40$, PSNR = 16.05 dB

(e) PSNR=29.86 dB (f) PSNR=26.11 dB (g) PSNR=23.49 dB (h) PSNR=22.84 dB

Fig. 5. Illustration of the denoising results.

5 Conclusion and Future Work

In this paper, we introduced a framework of learning functions from part of the data. We gave a data-dependent regularization problem which helps us learn a function using the manifold structure of the data. We used two applications to illustrate the learning framework. While these two applications are just part of the learning framework. They are special cases of the data-dependent regularization problem. However, for the general application, we need to calculate $||f_1||^2_{\mathcal{M}}$ and $||f_2||^2_{\mathcal{M}^c}$, which is hard to do so since we only have partial data. So we need to approximate $||f_1||^2_{\mathcal{M}}$ and $||f_2||^2_{\mathcal{M}^c}$ from incomplete data and to propose a new learning algorithm so that our framework can be used in a general application. This is part of our future work. Another line for the future work is from the theoretical aspect. We showed that the solution of the data-dependent regularization problem is the linear combination of the kernel. It then can be viewed as a function approximation result. If it is an approximated function, then we can consider the error analysis of the approximated function.

References

1. Schölkopf, B., Smola, A. J. and Bach, F.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, (2002).
2. Aronszajn, N.: Theory of reproducing kernels. Transactions of the American mathematical society, **68**(3), 337–404 (1950).
3. Schölkopf, B., Herbrich, R. and Smola, A. J.: A generalized representer theorem. In: International conference on computational learning theory, pp. 416–426. Springer, Berlin, Heidelberg (2001).
4. Argyriou, A., Micchelli, C. A. and Pontil, M.: When is there a representer theorem? Vector versus matrix regularizers. Journal of Machine Learning Research, **10**(Nov), 2507–2529 (2009).
5. Gohberg, I., Goldberg, S. and Kaashoek, M.A.: Hilbert-Schmidt Operators. In: Classes of Linear Operators, Vol. I, pp. 138–147,. Birkhäuser, Basel (1990).
6. Helmberg, G.: Introduction to spectral theory in Hilbert space. Courier Dover Publications, (2008).
7. Mikhail, B., Partha, N. and Vikas, S.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, **7**(Nov), 2507–2529 (2006).
8. Pipkin, A. C.: A course on integral equations (No. 9). Springer Science & Business Media, (1991).
9. Tikhonov A. N.: Regularization of incorrectly posed problems. Soviet Mathematics Doklady, **4**(6), 1624–1627 (1963).
10. Saitoh, S. and Sawano, Y.: Theory of reproducing kernels and applications. Singapore: Springer Singapore, (2016).
11. Rudin W.: Functional Analysis. MA: McGraw-Hill, Boston (1991).
12. Amir A. D., Luis G. C. R., Yukawa, M. and Stanczak, S.: Adaptive Learning for Symbol Detection: A Reproducing Kernel Hilbert Space Approach. Machine Learning for Future Wireless Communications, 197–211 (2020).
13. Kernel Functions for Machine Learning Applications, http://crsouza.com/2010/03/17/kernel-functions-for-machine-learning-applications/.