

Learning over Compact Metric Spaces

Ha Quang Minh¹ and Thomas Hofmann²

¹ Department of Mathematics
Brown University, Providence RI 02912-1917 USA
minh@math.brown.edu

² Department of Computer Science
Brown University, Providence RI 02912-1910, USA
th@cs.brown.edu

Abstract. We consider the problem of learning on a compact metric space X in a functional analytic framework. For a dense subalgebra of $Lip(X)$, the space of all Lipschitz functions on X , the Representer Theorem is derived. We obtain exact solutions in the case of least square minimization and regularization and suggest an approximate solution for the Lipschitz classifier.

1 Introduction

One important direction of current machine learning research is the generalization of the Support Vector Machine paradigm to handle the case where the input space is an arbitrary metric space. One such generalization method was suggested recently in [2], [5]: we embed the input space X into a Banach space E and the hypothesis space of decisions functions on X into the dual space E^* of linear functionals on E . In [5], the hypothesis space is $Lip(X)$, the space of all bounded Lipschitz functions on X . The input space X itself is embedded into the space $AE(X_0)$ of molecules on X_0 , which up to isometry, is the largest Banach space that X embeds into isometrically [6].

The Representer Theorem, which is essential in the formulation of the solutions of Support Vector Machines, was, however, not achieved in [2]. In order to obtain this theorem, it is necessary to restrict ourselves to subspaces of $Lip(X)$ consisting of functions of a given explicit form. In this paper, we introduce a general method for deriving the Representer Theorem and apply it to a dense subalgebra of $Lip(X)$. We then use the theorem to solve a problem of least square minimization and regularization on the subalgebra under consideration. Our approach can be considered as a generalization of the Lagrange polynomial interpolation formulation. It is substantially different from that in [5], which gives solutions that are minimal Lipschitz extensions (section 6.1).

Throughout the paper, (X, d) will denote a compact metric space and $S = \{(x_i, y_i)\}_{i=1}^n \subset (X \times \mathbb{R})^n$ a sample of length n .

1.1 The Representer Theorem

The Representer Theorem is not magic, neither is it an exclusive property of Support Vector Machines and Reproducing Kernel Hilbert Spaces. It is a direct

consequence of the fact that our training data is finite. A general method to derive the Representer Theorem is as follows. Let \mathcal{F} be a normed space of real-valued functions on X . Consider the evaluation operator

$$A_S : \mathcal{F} \rightarrow \mathbb{R}^n \quad (1)$$

defined by

$$A_S(f) = (f(x_1), \dots, f(x_n)) \quad (2)$$

Consider the problem of minimizing the following functional over \mathcal{F} :

$$I_S(f) = \sum_{i=1}^n V(f(x_i), y_i) \quad (3)$$

with V being a convex, lower semicontinuous loss function. Let $\ker(A_S)$ denote the kernel of the map A_S , defined by

$$\ker(A_S) = \{f \in \mathcal{F} : A_S(f) = (f(x_1), \dots, f(x_n)) = (0, \dots, 0)\} \quad (4)$$

Clearly, the problem of minimizing I_S over \mathcal{F} is equivalent to minimizing I_S over the quotient space $\mathcal{F}/\ker(A_S)$, which being isomorphic to the image $\text{Im}(A_S) \subset \mathbb{R}^n$, is finite dimensional. Let \mathcal{F}_n be the complementary subspace of $\ker(A_S)$

$$\mathcal{F} = \mathcal{F}_n \oplus \ker(A_S) \quad (5)$$

that is a linear subspace of \mathcal{F} such that $\mathcal{F}_n \cap \ker(A_S) = \{0\}$ and every $f \in \mathcal{F}$ admits a unique decomposition

$$f = f_n + r \quad (6)$$

where $f_n \in \mathcal{F}_n$ and $r \in \ker(A_S)$. Clearly we have $f - f_n \in \ker(A_S)$. Consider the equivalent relation on the quotient space $\mathcal{F}/\ker(A_S)$ defined by

$$f \sim f_0 \iff f \in [f_0] \iff A_S f = A_S f_0 \iff f - f_0 \in \ker(A_S) \quad (7)$$

Thus $f \sim f_0$ iff they have the same projection onto \mathcal{F}_n . Hence $\mathcal{F}/\ker(A_S) \cong \mathcal{F}_n$ via the identification.

$$[f] \rightarrow f_n \quad (8)$$

We are led to the following fundamental result:

Theorem 1. *There is **always** a minimizer of I_S , if one exists, lying in a finite dimensional subspace \mathcal{F}_n of \mathcal{F} , with dimension at most n . The space \mathcal{F}_n is the complementary subspace of $\ker(A_S)$.*

Proof. From the preceding discussion, it clearly follows that the problem of minimizing I_S over \mathcal{F} is equivalent to minimizing I_S over the subspace \mathcal{F}_n . This subspace has dimension at most n

$$\dim(\mathcal{F}_n) = \dim(\mathcal{F}/\ker(A_S)) \leq n$$

Thus if I_S has minimizers in \mathcal{F} , then it must have one minimizer lying in \mathcal{F}_n . \square

Corollary 1. *Suppose the problem of minimizing I_S over \mathcal{F}_n has a set of solutions F^* , then the set of all minimizers of I_S over \mathcal{F} has the form*

$$F^* + \ker(A_S) = \{f^* + r \mid f^* \in F^*, r \in \ker(A_S)\} \quad (9)$$

Proof. This is obvious. \square

Consider now the problem of minimizing the regularized functional

$$I_{S,\gamma}(f) = \sum_{i=1}^n V(f(x_i), y_i) + \gamma\Omega(f) \quad (10)$$

where Ω is a strictly convex, coercive functional on \mathcal{F} . We have another key result:

Theorem 2. *The functional $I_{S,\gamma}$ has a unique minimizer in \mathcal{F} . Assume further that the regularizer Ω satisfies:*

$$\Omega(f) = \Omega(f_n + r) \geq \Omega(f_n) \quad (11)$$

for all $f \in \mathcal{F}$, where $f_n \in \mathcal{F}_n$ and $r \in \ker(A_S)$. Then this minimizer lies in the finite dimensional subspace \mathcal{F}_n .

Proof. The existence and uniqueness of the minimizer f^* is guaranteed by the coercivity and strict convexity of the regularizer Ω , respectively. If furthermore, $\Omega(f_n + r) \geq \Omega(f_n)$ then for all $f \in \mathcal{F}$:

$$I_{S,\gamma}(f) \geq I_{S,\gamma}(f_n)$$

Thus a function f^* minimizing $I_{S,\gamma}$ must lie in the finite dimensional subspace \mathcal{F}_n of \mathcal{F} . \square

Without the assumption of strict convexity and coercivity of the functional Ω , we can no longer state the uniqueness or existence of the minimizer, but we still have the following result

Theorem 3. *Suppose the functional Ω satisfies*

$$\Omega(f) = \Omega(f_n + r) \geq \Omega(f_n) \quad (12)$$

for all $f \in \mathcal{F}$, where $f_n \in \mathcal{F}_n$ and $r \in \ker(A_S)$, with equality iff $r = 0$. If the problem of minimizing $I_{S,\gamma}$ over \mathcal{F} has a solution f^* , it must lie in the finite dimensional subspace \mathcal{F}_n .

Proof. This is similar to the above theorem. \square

Having the above key results, the Representer Theorem can then be obtained if we can exhibit a basis for the above finite dimensional subspace \mathcal{F}_n via the data points x_i ($1 \leq i \leq n$).

Example 1 (RKHS). Let $\mathcal{F} = \mathcal{H}_K$ be the reproducing kernel Hilbert space induced by a Mercer kernel K , then from the reproducing property $f(x) = \langle f, K(x, \cdot) \rangle$, it follows that

$$\ker(A_S) = \text{span}\{K(x_i, \cdot)_{i=1}^n\}^\perp$$

From the unique orthogonal decomposition

$$\mathcal{H}_K = \text{span}\{K(x_i, \cdot)_{i=1}^n\} \oplus \text{span}\{K(x_i, \cdot)_{i=1}^n\}^\perp$$

it follows that $\mathcal{F}_n = \text{span}\{K(x_i, \cdot)_{i=1}^n\}$. \square

In section 2, we apply the above framework to derive the Representer Theorem for the special case \mathcal{F} is the vector space of all algebraic polynomials on a compact subset of the real line \mathbb{R} . We then generalize this result to the case of a general compact metric space in sections 3 and 4.

2 Learning over Compact Subsets of \mathbb{R}

Let $X \subset \mathbb{R}$ be compact. Let $P(X)$ be the vector space of all algebraic polynomials on X , then $P(X)$ is dense in $C(X)$ according to Weierstrass Approximation Theorem:

Theorem 4 (Weierstrass Approximation Theorem). *Each continuous function $f \in C(X)$ is uniformly approximable by algebraic polynomials: for each $\epsilon > 0$, there is a polynomial $p \in P(X)$ such that*

$$|f(x) - p(x)| < \epsilon \tag{13}$$

for all $x \in X$.

Consider the problem of minimizing the functional I_S over $P(X)$.

Lemma 1.

$$\ker(A_S) = \{f \in P(X) : f(x) = (x - x_1) \dots (x - x_n)r_n(x)\} \tag{14}$$

for some $r_n \in P(X)$. Let $P_n(X) = \text{span}\{1, (x - x_1), (x - x_1)(x - x_2), \dots, (x - x_1) \dots (x - x_{n-1})\}$, then $P(X)$ admits the following unique decomposition

$$P(X) = P_n(X) \oplus \ker(A_S) \tag{15}$$

Proof. First we note that $f \in \ker(A_S) \iff (f(x_1), \dots, f(x_n)) = (0, \dots, 0)$ iff x_i ($1 \leq i \leq n$) is a zero of f iff f contains the linear factor $(x - x_i)$ ($1 \leq i \leq n$), hence the form of $\ker(A_S)$.

To prove the unique decomposition, we apply the Taylor expansion to f , with centers x_1, \dots, x_n successively:

$$\begin{aligned} f(x) &= c_0 + (x - x_1)r_1(x) = c_0 + (x - x_1)[c_1 + (x - x_2)r_2(x)] \\ &= c_0 + c_1(x - x_1) + (x - x_1)(x - x_2)r_2(x) = \dots = \\ &= c_0 + c_1(x - x_1) + c_2(x - x_1)(x - x_2) + \dots + c_{n-1}(x - x_1) \dots (x - x_{n-1}) \\ &\quad + (x - x_1) \dots (x - x_n)r_n(x) \end{aligned}$$

with $c_i \in \mathbb{R}$ ($0 \leq i \leq n-1$). □

The basis $\{\prod_{j=1}^i (x-x_j)\}_{i=0}^{n-1}$ for $P_n(X)$ is not symmetric in the x_i 's. Let us construct a symmetric basis for this subspace.

Lemma 2.

$$P_n(X) = \text{span}\left\{\prod_{\substack{j=1 \\ j \neq i}}^n (x-x_j)\right\}_{i=1}^n \quad (16)$$

Proof. Let $f = \sum_{i=0}^{n-1} c_i \prod_{j=1}^i (x-x_j)$. Define the function

$$g^*(x) = \sum_{i=1}^n d_i \prod_{\substack{j=1 \\ j \neq i}}^n (x-x_j) \quad (17)$$

with

$$d_i = \frac{\sum_{j=0}^{i-1} c_j \prod_{k=1}^j (x_i - x_k)}{\prod_{j \neq i} (x_i - x_j)} \quad (18)$$

It is straightforward to verify that $f^*(x_i) = g^*(x_i)$ ($1 \leq i \leq n$). Since f^* and g^* have degree $n-1$, it follows that $f^* = g^*$. □

We arrive at the following Representer Theorem for the space $P(X)$:

Theorem 5 (Representer Theorem). *The problem of minimizing the functional I_S over space $P(X)$ is equivalent to minimizing I_S over the finite-dimensional subspace $P_n(X) = \text{span}\{\prod_{j \neq i} (x-x_j)\}_{i=1}^n$. Suppose the latter problem has a set of solutions F^* , then the set of all minimizers of $I_S(f)$ over $P(X)$ has the form:*

$$F^* + \ker(A_S) = \{f^* + (x-x_1) \dots (x-x_n)r_n \mid f^* \in F^*, r_n \in P(X)\} \quad (19)$$

Each $f^* \in F^*$ admits a unique representation:

$$f^* = \sum_{i=1}^n c_i \prod_{\substack{j=1 \\ j \neq i}}^n (x-x_j) \quad (20)$$

for $c_i \in \mathbb{R}$ ($1 \leq i \leq n$).

Proof. This is a special case of theorem 1, with $\mathcal{F}_n = P_n(X)$. □

3 The Stone-Weierstrass Theorem

Let us now consider the general case where X is a compact metric space. We then have Stone's generalization of Weierstrass Approximation Theorem. For a very accessible treatment of this topic, we refer to [1].

Definition 1 (Algebra). *A real algebra is a vector space \mathcal{A} over \mathbb{R} together with a binary operation representing multiplication: $x, y \in \mathcal{A} \rightarrow xy \in \mathcal{A}$ satisfying:*

(i) *Bilinearity: for all $a, b \in \mathbb{R}$ and all $x, y, z \in \mathcal{A}$:*

$$(a.x + b.y)z = a.xz + b.yz$$

$$x(a.y + b.z) = a.xy + b.xz$$

(ii) *Associativity:* $x(yz) = (xy)z$

The multiplicative identity, if it exists, is called the unit of the algebra. An algebra with unit is called a unital algebra. A complex algebra over \mathbb{C} is defined similarly.

Definition 2 (Normed algebra-Banach algebra). A normed algebra is a pair $(\mathcal{A}, \|\cdot\|)$ consisting of an algebra \mathcal{A} together with a norm $\|\cdot\| : \mathcal{A} \rightarrow [0, \infty)$ satisfying

$$\|xy\| \leq \|x\| \|y\| \tag{21}$$

A Banach algebra is a normed algebra that is a Banach space relative to its given norm.

Example 2. $C(X)$: Let X be a compact Hausdorff space. We have the unital algebra $C(X)$ of all real-valued functions on X , with multiplication and addition being defined pointwise:

$$fg(x) = f(x)g(x) \text{ and } (f + g)(x) = f(x) + g(x)$$

Relative to the supremum norm $\|\cdot\|_\infty$, $C(X)$ is a commutative Banach algebra with unit.

Definition 3 (Separation). Let X be a metric space. Let \mathcal{A} be a set of real-valued functions on X . \mathcal{A} is said to separate the points of X if for each pair x, y of distinct points of X there exists a function $f \in \mathcal{A}$ such that $f(x) \neq f(y)$.

Theorem 6 (Stone-Weierstrass Theorem). Let X be a compact metric space and \mathcal{A} a subalgebra of $C(X)$ that contains the constant functions and separates the points of X . Then \mathcal{A} is dense in the Banach space $C(X)$.

4 Learning over Compact Metric Spaces

Let (X, d) be a compact metric space containing at least two points.

Proposition 1. Let \mathcal{A} be the subalgebra of $C(X)$ generated by the family

$$\{1, \phi_x : t \rightarrow d(x, t)\}_{x \in X} \tag{22}$$

where 1 denote the constant function with value 1, then \mathcal{A} is dense in $C(X)$.

Proof. By the Stone-Weierstrass Theorem, we need to verify that \mathcal{A} separates the points of X . Let t_1, t_2 be two distinct points in X , so that $d(t_1, t_2) \neq 0$. Suppose that $d(x, t_1) = d(x, t_2)$ for all $x \in X$. Let $x = t_1$, we then obtain:

$$d(t_1, t_2) = d(t_1, t_1) = 0$$

a contradiction. Thus there must exist $x \in X$ such that $d(x, t_1) \neq d(x, t_2)$, showing that \mathcal{A} separates the points in X . \square

Consider the algebra \mathcal{A} defined in the above proposition and the problem of minimizing I_S over \mathcal{A} .

Lemma 3. *Each $f \in \mathcal{A}$ can be expressed in the form:*

$$f = g + d(x_1, \cdot) \dots d(x_n, \cdot) f_{n+1} \quad (23)$$

where

$$g = f_1 + d(x_1, \cdot) f_2 + d(x_1, \cdot) d(x_2, \cdot) f_3 + \dots + d(x_1, \cdot) d(x_2, \cdot) \dots d(x_{n-1}, \cdot) f_n \quad (24)$$

and $f_{n+1} \in \mathcal{A}$, $f_i \in \mathcal{A}/\langle d(x_i, \cdot) \rangle$ with $\langle d(x_i, \cdot) \rangle$ being the ideal generated by $d(x_i, \cdot)$, $1 \leq i \leq n$.

Proof. This is similar to a Taylor expansion: clearly there is $f_i \in \mathcal{A}/\langle d(x_i, \cdot) \rangle$ such that

$$\begin{aligned} f &= f_1 + d(x_1, \cdot) r_1 = f_1 + d(x_1, \cdot) [f_2 + d(x_2, \cdot) r_2] \\ &= f_1 + d(x_1, \cdot) f_2 + d(x_1, \cdot) d(x_2, \cdot) r_2 \end{aligned}$$

Continuing in this way we obtain the lemma. \square

Since $f(x_i) = g(x_i)$ ($1 \leq i \leq n$), minimizing I_S over \mathcal{A} is equivalent to minimizing I_S over all f of the form:

$$f = f_1 + d(x_1, \cdot) f_2 + d(x_1, \cdot) d(x_2, \cdot) f_3 + \dots + d(x_1, \cdot) d(x_2, \cdot) \dots d(x_{n-1}, \cdot) f_n \quad (25)$$

with $f_i \in \mathcal{A}/\langle d(x_i, \cdot) \rangle$. From the above equation, we obtain for $1 \leq i \leq n$:

$$f(x_i) = \sum_{j=1}^i f_j(x_i) \prod_{k=1}^{j-1} d(x_k, x_i) \quad (26)$$

It is straightforward to verify that

$$f = \sum_{k=1}^n \sum_{i=k}^n f_k(x_i) \frac{\prod_{j \neq i} d(x_j, \cdot)}{\prod_{j=k, j \neq i}^n d(x_j, x_i)} \quad (27)$$

From the above general expression for f , it follows that there are constants $c_i \in \mathbb{R}$ ($1 \leq i \leq n$) such that

$$f = \sum_{i=1}^n c_i \prod_{j \neq i} d(x_j, \cdot) \quad (28)$$

Let $P_n(X)$ denote the n -dimensional subspace of \mathcal{A} defined by

$$P_n(X) = \text{span} \left\{ \prod_{j \neq i} d(x_j, \cdot) \right\}_{i=1}^n \quad (29)$$

We have proved the following theorem:

Theorem 7 (Representer Theorem). *The problem of minimizing the functional I_S over \mathcal{A} is equivalent to minimizing I_S over the n -dimensional subspace $P_n(X) = \text{span}\{\prod_{j \neq i} d(x_j, \cdot)\}_{i=1}^n$. Suppose the latter problem has a set of solutions F^* , then the set of minimizer of I_S over \mathcal{A} has the form*

$$\{f^* + d(x_1, \cdot) \dots d(x_n, \cdot) f_{n+1} : f^* \in F^*, f_{n+1} \in \mathcal{A}\} \quad (30)$$

Each $f^* \in F^*$ admits a unique representation

$$f^* = \sum_{i=1}^n c_i \prod_{j \neq i} d(x_j, \cdot) \quad (31)$$

for $c_i \in \mathbb{R}$ ($1 \leq i \leq n$). Let Ω be as in theorem 2, then the problem of minimizing the functional $I_{S,\gamma}$ over \mathcal{A} has a unique solution lying in $P_n(X)$.

Proof. This is a special case of theorems 1 and 2, with $\mathcal{F}_n = P_n(X)$. \square

We now show that the algebra \mathcal{A} consists of Lipschitz functions and that it is dense in the space $Lip(X)$ of all Lipschitz functions on X , in the supremum norm:

Lemma 4. *For each $x \in X$, the function $\phi_x : t \rightarrow d(x, t)$ is Lipschitz with Lipschitz constant $L(\phi_x) = 1$.*

Proof. Let $t_1, t_2 \in X$. From the triangle inequality, we have:

$$d(x, t_1) \leq d(x, t_2) + d(t_2, t_1) \Rightarrow d(x, t_1) - d(x, t_2) \leq d(t_1, t_2)$$

Similarly, we have $d(x, t_2) - d(x, t_1) \leq d(t_1, t_2)$. It follows that

$$|\phi_x(t_1) - \phi_x(t_2)| = |d(x, t_1) - d(x, t_2)| \leq d(t_1, t_2)$$

with equality iff $t_1 = x$ or $t_2 = x$. Thus ϕ_x is a Lipschitz function with Lipschitz constant $L(\phi_x) = 1$. \square

Proposition 2. *Let X be a compact metric space and \mathcal{A} defined as above. Then \mathcal{A} consists of Lipschitz functions and \mathcal{A} is dense in $Lip(X)$ in the supremum norm.*

Proof. Since Lipschitz functions are closed under addition, scalar multiplication, and for X bounded, pointwise multiplication (see appendix), it follows from the above lemma that \mathcal{A} consists of Lipschitz functions, that is \mathcal{A} is a subalgebra of $Lip(X)$. Since for compact X , both \mathcal{A} and $Lip(X)$ are dense in $C(X)$ in the supremum norm, it follows that \mathcal{A} is dense in $Lip(X)$ in the supremum norm. \square

5 Least Square Minimization and Regularization

5.1 Least Square Minimization

Let $S = \{(x_i, y_i)\}_{i=1}^n \in (X \times \mathbb{R})^n$ be a training sample of length n . Consider the problem of minimizing the empirical square error over \mathcal{A} :

$$I_S(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 \quad (32)$$

or equivalently

$$I_S(f) = \sum_{i=1}^n (f(x_i) - y_i)^2 \quad (33)$$

By the Representer Theorem, this is equivalent to minimizing the functional $I_S(f)$ over the finite dimensional subspace $P_n(X)$. Let $f = \sum_{i=1}^n c_i \prod_{j \neq i} d(x_j, \cdot) \in Lip(X)$. Let

$$M_i = \prod_{j \neq i} d(x_j, x_i)$$

then clearly

$$f(x_i) = c_i M_i$$

Theorem 8. *The problem of minimizing the functional $I_S(f)$ over the finite dimensional subspace $P_n(X)$ has a unique solution*

$$f^* = \sum_{i=1}^n \frac{y_i}{M_i} \prod_{j \neq i} d(x_j, \cdot) = \sum_{i=1}^n y_i \frac{\prod_{j \neq i} d(x_j, \cdot)}{\prod_{j \neq i} d(x_j, x_i)} \quad (34)$$

Proof. Each $f \in P_n(X)$ has the form: $f = \sum_{i=1}^n c_i \prod_{j \neq i} d(x_j, \cdot)$. Thus

$$f(x_i) = c_i \prod_{j \neq i} d(x_j, x_i) = c_i M_i$$

Clearly the smallest value that $I_S(f)$ assumes is zero, which occurs iff

$$f(x_i) = y_i \iff c_i M_i = y_i \iff c_i = \frac{y_i}{M_i}$$

This gives us the desired minimizer f^* . □

Remark 1. Let $\phi_i(x) = \frac{\prod_{j \neq i} d(x_j, \cdot)}{\prod_{j \neq i} d(x_j, x_i)}$, then we have

$$\phi_i(x_j) = \delta_{ij} \text{ and } f^*(x) = \sum_{i=1}^n y_i \phi_i(x)$$

In the case $X \subset \mathbb{R}$, these functions are precisely the Lagrange interpolation polynomials and we recover the Lagrange interpolation formula.

5.2 Least Square Regularization

The minimization process above always gives an exact interpolation, which may lead to the undesirable phenomenon of overfitting. Hence we consider the following regularization problem. Each function $f \in \mathcal{A}$ has the form $f = \sum_{J \subset I} c_J \prod_{j \in J} d(x_j, \cdot)$ where I is a finite index set. Consider the functional $\Omega : \mathcal{A} \rightarrow \mathbb{R}$ defined by

$$\Omega(f) = \sum_{J \subset I} |c_J|^2 \quad (35)$$

Lemma 5. *Let $f \in \mathcal{A}$ with the decomposition: $f = g + d(x_1, \cdot) \dots d(x_n, \cdot) f_{n+1}$ where $g \in P_n(X)$ and $f_{n+1} \in \mathcal{A}$. Then $\Omega(f) = \Omega(g) + \Omega(f_{n+1})$.*

Proof. This is obvious. \square

Lemma 6. *The functional Ω is strictly convex.*

Proof. This follows from the strict convexity of the square function. \square

Lemma 7. *Let $f = \sum_{J \subset I} c_J \prod_{j \in J} d(x_j, \cdot) \in \mathcal{A}$. Then*

$$\|f\|_\infty \leq \sum_{J \subset I} |c_J| \text{diam}(X)^{|J|} \leq \left(\sum_{J \subset I} \text{diam}(X)^{2|J|} \right)^{1/2} \left(\sum_{J \subset I} |c_J|^2 \right)^{1/2} \quad (36)$$

The functional Ω is coercive in the supremum norm:

$$\lim_{\|f\|_\infty \rightarrow \infty} \Omega(f) = \infty \quad (37)$$

Proof. We have

$$\| \prod_{j \in J} d(x_j, \cdot) \|_\infty \leq \text{diam}(X)^{|J|}$$

It follows that

$$\|f\|_\infty \leq \sum_{J \subset I} |c_J| \text{diam}(X)^{|J|} \leq \left(\sum_{J \subset I} \text{diam}(X)^{2|J|} \right)^{1/2} \left(\sum_{J \subset I} |c_J|^2 \right)^{1/2}$$

Thus $\|f\|_\infty \rightarrow \infty$ implies that $\sum_{J \subset I} |c_J|^2 \rightarrow \infty$ as well, showing that Ω is coercive in the supremum norm. \square

Lemma 8. *Let $f = d(x_1, \cdot) \dots d(x_k, \cdot)$. Then*

$$L(f) \leq k \text{diam}(X)^{k-1} \quad (38)$$

Let $f = \sum_{J \subset I} c_J \prod_{j \in J} d(x_j, \cdot)$. Then there is a constant $C > 0$ such that

$$L(f) \leq C \sum_{J \subset I} |c_J| \leq C \left(\sum_{J \subset I} 1 \right)^{1/2} \left(\sum_{J \subset I} |c_J|^2 \right)^{1/2} \quad (39)$$

In particular, for $f = \sum_{i=1}^n c_i \prod_{j \neq i} d(x_j, \cdot)$, we have

$$L(f) \leq C \sum_{i=1}^n |c_i| \leq C \sqrt{n} \left(\sum_{i=1}^n |c_i|^2 \right)^{1/2} \quad (40)$$

Proof. The first inequality follows from a standard induction argument. This and the Cauchy-Schwarz inequality imply the other inequalities. \square

Consider the problem of minimizing the regularized functional:

$$I_{S,\gamma}(f) = \sum_{i=1}^n (f(x_i) - y_i)^2 + \gamma \Omega(f) \quad (41)$$

with regularization parameter $\gamma > 0$. By lemmas 7 and 8, this regularization process aims to minimize $\sum_{i=1}^n (f(x_i) - y_i)^2$ and penalize $\|f\|_\infty$ and $L(f)$ simultaneously.

Theorem 9. *The problem of minimizing the regularized functional $I_{S,\gamma}(f)$ over the algebra \mathcal{A} has a unique solution f^* which lies in the finite dimensional subspace $P_n(X)$:*

$$f^* = \sum_{i=1}^n \frac{y_i M_i}{\gamma + M_i^2} \prod_{j \neq i} d(x_j, \cdot) \quad (42)$$

Proof. The functional Ω is strictly convex and coercive in the supremum norm on \mathcal{A} and satisfies $\Omega(f) = \Omega(g) + \Omega(f_{n+1}) \geq \Omega(g)$. Thus by the Representer Theorem, there is a unique solution minimizing $I_{S,\gamma}(f)$, which lies in the finite dimensional subspace $P_n(X)$. We have for $f \in P_n(X)$:

$$I_{S,\gamma}(f) = \sum_{i=1}^n [(c_i M_i - y_i)^2 + \gamma c_i^2] = \sum_{i=1}^n [c_i^2 (\gamma + M_i^2) - 2c_i M_i y_i + y_i^2]$$

Differentiating and setting $\frac{\partial I}{\partial c_i} = 2c_i (\gamma + M_i^2) - 2M_i y_i = 0$, we obtain

$$c_i = \frac{y_i M_i}{\gamma + M_i^2}$$

as claimed. \square

6 The Lipschitz Classifier

Let $(x_i, y_i)_{i=1}^n \subset (X \times \{\pm 1\})^n$ be a set of training data, with the assumption that both classes ± 1 are present. Let $X_0 = X \cup \{e\}$ where e is a distinguished base point with the metric $d^{X_0}|_{X \times X} = d$ and $d^{X_0}(x, e) = \text{diam}(X)$ for $x \in X$. It is straightforward to show that

Proposition 3 ([5]). *$Lip(X)$ is isometrically isomorphic to $Lip_0(X_0)$ via the map $\psi : Lip(X) \rightarrow Lip_0(X_0)$ defined by $(\psi f)(x) = f(x)$ for $x \in X$ and $(\psi f)(e) = 0$. One has $\|f\|_L = L(\psi f)$.*

Proposition 4 ([6]). *X embeds isometrically into the Banach space $AE(X_0)$, via the map $\Phi(x) = m_x = m_{xe} = \chi_x - \chi_e$. $Lip_0(X_0)$ embeds isometrically into the dual space $AE(X_0)^*$, via the map $T : Lip_0(X_0) \rightarrow AE(X_0)^*$ defined by $\langle Tf, m \rangle = \sum_{x \in X} f(x)m(x)$ for all $f \in Lip_0(X_0)$, all $m \in AE(X_0)$. Clearly $f(x) = \langle Tf, m_x \rangle$ for all $x \in X$.*

The problem of finding a decision function $f \in Lip(X)$ separating the points x_i 's in X is then equivalent to that of finding the corresponding linear functional $Tf \in AE(X_0)^*$ separating the corresponding molecules m_{x_i} , that is a hyperplane Hf defined by $Hf = \{m \in AE(X_0) : \langle Tf, m \rangle = 0\}$. It is straightforward to show the following

Proposition 5 (Margin of the Lipschitz Classifier [5]). *Assume that the hyperplane is normalized such that $\min_{1 \leq i \leq n} |f(x_i)| = 1$ and suppose that $y_i f(x_i) \geq 1$ ($1 \leq i \leq n$). Then*

$$\rho = \inf_{1 \leq i \leq n, m_h \in Hf} \|m_{x_i} - m_h\|_{AE} \geq \frac{1}{L(f)} \quad (43)$$

Thus the following algorithm then corresponds to a large margin algorithm in the space $AE(X_0)$:

Algorithm 1 ([5])

$$\text{Minimize}_{f \in Lip(X)} L(f) \quad \text{subject to } y_i f(x_i) \geq 1 \quad (1 \leq i \leq n) \quad (44)$$

The solutions of this algorithm are precisely the minimal Lipschitz extensions of the function $f : \{x_i\}_{i=1}^n \rightarrow \{\pm 1\}$ with $f(x_i) = y_i$, as we show below.

6.1 Minimal Lipschitz Extensions

The following was shown simultaneously in 1934 by McShane [4] and Whitney [7].

Proposition 6 (Minimal Lipschitz Extension-MLE). *Let (X, d) denote an arbitrary metric space and let E be any nonempty subset of X . Let $f : E \rightarrow \mathbb{R}$ be a Lipschitz function. Then there exists a **minimal Lipschitz extension** of f to X , that is a Lipschitz function $h : X \rightarrow \mathbb{R}$ such that $h|_E = f$ and $L(h) = L(f)$.*

Proof. Two such minimal Lipschitz extensions were constructed explicitly in [4] and [7]:

$$\overline{f}(x) = \inf_{y \in E} \{f(y) + L(f)d(x, y)\} \quad (45)$$

$$\underline{f}(x) = \sup_{y \in E} \{f(y) - L(f)d(x, y)\} \quad (46)$$

Furthermore, if u is any minimal Lipschitz extension of f to X , then for all $x \in X$:

$$\underline{f}(x) \leq u(x) \leq \overline{f}(x) \quad (47)$$

We refer to the above references for detail. \square

Let us return to the classification problem. Let $E = \{x_i\}_{i=1}^n$ and $f : E \rightarrow \{\pm 1\}$ be defined by $f(x_i) = y_i$. Let X^+ and X^- denote the sets of training points with positive and negative labels, respectively. Let $d(X^+, X^-) = \inf_{x \in X^+, x' \in X^-} d(x, x')$. It is straightforward to see that f is Lipschitz with Lipschitz constant $L^* = \frac{2}{d(X^+, X^-)}$. The above proposition gives two of f 's minimal Lipschitz extensions:

$$\bar{f}(x) = \min_i \{y_i + L^* d(x, x_i)\} \text{ and } \underline{f}(x) = \max_i \{y_i - L^* d(x, x_i)\}$$

These are precisely the solutions of the above algorithm in [5].

Remark 2. The notion of minimal Lipschitz extension is not completely satisfactory. Firstly, it is not unique. Secondly, and more importantly, it involves only the global Lipschitz constant and ignores what may happen locally. For a discussion of this phenomenon, we refer to [3].

6.2 A Variant of the Lipschitz Classifier

The problem of computing the Lipschitz constants for a class of functions is nontrivial in general. It is easier to obtain an upper bound for $L(f)$ and minimize it instead. Let us consider this approach with the algebra \mathcal{A} , which is dense in $Lip(X)$ in the supremum norm as shown above.

From the above upper bound on $L(f)$, instead of minimizing $L(f)$, we can minimize $\sum_{J \subset I} |c_J|$. We obtain the following algorithm:

Algorithm 2

$$\text{Minimize}_{I \subset \mathbb{N}} \sum_{J \subset I} |c_J| \text{ subject to } y_i f(x_i) \geq 1 \text{ (} 1 \leq i \leq n \text{)} \quad (48)$$

The functional $\Omega : \mathcal{A} \rightarrow \mathbb{R}$ defined by

$$\Omega(f) = \sum_{J \subset I} |c_J| \quad (49)$$

clearly satisfies $\Omega(g + d(x_1, \cdot) \dots d(x_n, \cdot) f_{n+1}) \geq \Omega(g)$ for all $g \in P_n(X)$ and $f_{n+1} \in \mathcal{A}$, with equality iff $f_{n+1} = 0$. Thus by theorem 3, we have the equivalent problem:

Algorithm 3

$$\text{Minimize} \sum_{i=1}^n |c_i| \text{ subject to } y_i c_i M_i \geq 1 \text{ (} 1 \leq i \leq n \text{)} \quad (50)$$

According to lemma 7, the functional Ω is coercive in the $\|\cdot\|_\infty$ norm, thus the problem has a solution. Let us show that it is unique and find its explicit form.

Theorem 10. *The above minimization problem has a unique solution*

$$f^* = \sum_{i=1}^n \frac{y_i}{M_i} \prod_{j \neq i} d(x_j, \cdot) = \sum_{i=1}^n y_i \frac{\prod_{j \neq i} d(x_j, \cdot)}{\prod_{j \neq i} d(x_j, x_i)} \quad (51)$$

Proof. $\sum_{i=1}^n |c_i|$ is obviously minimum when $y_i c_i M_i = 1$, implying that

$$c_i = \frac{y_i}{M_i}$$

as we claimed. □

Remark 3. Clear we have $f(x_i) = y_i$. From lemma 8, we have $L(f) \leq C \sum_{i=1}^n |c_i|$. Thus it follows that

$$\rho \geq \frac{1}{L(f)} \geq \frac{1}{C \sum_{i=1}^n |c_i|}$$

Thus the above algorithm can also be viewed as a large margin algorithm as well.

7 Conclusion

We presented a general method for deriving the Representer Theorem in learning algorithms. The method is applied to a dense subalgebra of the space of Lipschitz functions on a general compact metric space X . We then used the Representer Theorem to obtain solutions to several special minimization and regularization problems. This approach may be used to obtain solutions when minimizing other functionals over other function spaces as well. We plan to continue with a more systematic regularization method and comprehensive analysis of our approach in future research.

A Lipschitz Functions and Lipschitz Spaces

We review some basic properties of Lipschitz functions and the corresponding function spaces. For detail treatment we refer to [6]. Let X be a metric space. A function $f : X \rightarrow \mathbb{R}$ (or \mathbb{C}) is called Lipschitz if there is a constant L such that for all $x, y \in X$:

$$|f(x) - f(y)| \leq Ld(x, y) \tag{52}$$

The smallest such L is called the Lipschitz constant of f , denoted by $L(f)$. We have

$$L(f) = \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)} \tag{53}$$

Proposition 7 ([6]). *Let X be a metric space and f, g, f_n ($n \in \mathbb{N}$) be Lipschitz functions from X into \mathbb{R} (or \mathbb{C}). Then:*

(a) $L(af) = |a|L(f)$ for all $a \in \mathbb{R}$

(b) $L(f + g) \leq L(f) + L(g)$

Proposition 8 ([6]). *Let X be a metric space and $f, g : X \rightarrow \mathbb{R}$ (\mathbb{C}) be bounded Lipschitz functions. Then*

(a) $L(fg) \leq \|f\|_\infty L(g) + \|g\|_\infty L(f)$

(b) *If $\text{diam}(X) < \infty$, then the product of any two scalar-valued Lipschitz functions is again Lipschitz.*

Definition 4 ([6]). Let X be a metric space. $Lip(X)$ is the space of all bounded Lipschitz functions on X equipped with the Lipschitz norm:

$$\|f\|_L = \max\{\|f\|_\infty, L(f)\}$$

If \mathcal{X} is a bounded metric space, that is $diam(X) < \infty$, we follow [5] and define:

$$\|f\|_L = \max\left\{\frac{\|f\|_\infty}{diam(X)}, L(f)\right\}$$

Theorem 11 ([6]). $Lip(X)$ is a Banach space. If X is compact, then $Lip(X)$ is dense in $C(X)$ in the supremum norm.

Definition 5. Let X_0 be a pointed metric space, with a distinguished base point e . Then we define

$$Lip_0(X_0) = \{f \in Lip(X_0) : f(e) = 0\} \quad (54)$$

On this space, $L(f)$ is a norm.

Definition 6 (Arens-Eells Space). Let X be a metric space. A **molecule** of X is a function $m : X \rightarrow \mathbb{R}$ (or \mathbb{C}) that is supported on a finite set of X and that satisfies:

$$\sum_{x \in \mathcal{X}} m(x) = 0$$

For $x, y \in X$, define the molecule $m_{xy} = \chi_x - \chi_y$, where χ_x and χ_y denote the characteristic functions of the singleton sets $\{x\}$ and $\{y\}$. On the set of molecules, consider the norm:

$$\|m\|_{AE} = \inf\left\{\sum_{i=1}^n |a_i| d(x_i, y_i) : m = \sum_{i=1}^n a_i m_{x_i y_i}\right\}$$

The Arens-Eells space $AE(X)$ is defined to be the completion of the space of molecules under the above norm.

References

1. D. Bridges, *Foundations of Real and Abstract Analysis*, Graduate Texts in Mathematics 174, Springer, New York, 1998.
2. M. Hein and O. Bousquet, Maximal Margin Classification for Metric Spaces, *Proceedings of the 16th Conference on Learning Theory (COLT 2003)*, Washington DC, August 2003.
3. P. Juutinen, Absolutely Minimizing Lipschitz Extensions on a Metric Space, *Annales Academiæ Scientiarum Fennicæ Mathematica*, vol. 27, pages 57-67, 2002.
4. E.J. McShane, Extension of Range of Functions, *Bulletin of the American Mathematical Society*, vol. 40, pages 837-842, 1934.
5. U. von Luxburg and O. Bousquet, Distance-Based Classification with Lipschitz Functions, *Proceedings of the 16th Conference on Learning Theory (COLT 2003)*, Washington DC, August 2003.
6. N. Weaver, *Lipschitz Algebras*, World Scientific, Singapore, 1999.
7. H. Whitney, Analytic Extensions of Differentiable Functions Defined in Closed Sets, *Transactions of the American Mathematical Society*, vol. 36, no. 1, pages 63-89, 1934.