Journal of Machine Learning Research 5 (2004) 669-695 Submitted 1/04; Published 6/04

Distance–Based Classification with Lipschitz Functions

Ulrike von Luxburg ULRIKE.LUXBURG@TUEBINGEN.MPG.DE
Olivier Bousquet OLIVIER.BOUSQUET@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
Spemannstrasse 38
72076 Tübingen, Germany

Editors: Kristin Bennett and Nicolò Cesa-Bianchi

Abstract

The goal of this article is to develop a framework for large margin classification in metric spaces. We want to find a generalization of linear decision functions for metric spaces and define a corresponding notion of margin such that the decision function separates the training points with a large margin. It will turn out that using Lipschitz functions as decision functions, the inverse of the Lipschitz constant can be interpreted as the size of a margin. In order to construct a clean mathematical setup we isometrically embed the given metric space into a Banach space and the space of Lipschitz functions into its dual space. To analyze the resulting algorithm, we prove several representer theorems. They state that there always exist solutions of the Lipschitz classifier which can be expressed in terms of distance functions to training points. We provide generalization bounds for Lipschitz classifiers in terms of the Rademacher complexities of some Lipschitz function classes. The generality of our approach can be seen from the fact that several well-known algorithms are special cases of the Lipschitz classifier, among them the support vector machine, the linear programming machine, and the 1-nearest neighbor classifier.

1. Introduction

Support vector machines (SVMs) construct linear decision boundaries in Hilbert spaces such that the training points are separated with a large margin. The goal of this article is to extend this approach from Hilbert spaces to metric spaces: we want to find a generalization of linear decision functions for metric spaces and define a corresponding notion of margin such that the decision function separates the training points with a large margin. The reason why we are interested in metric spaces is that in many applications it is easier or more natural to construct distance functions between objects in the data space than positive definite kernel functions as they are used for support vector machines. Examples for this situation are the edit distance used to compare strings or graphs and the earth mover's distance on images.

SVMs can be seen from two different points of view. In the regularization interpretation, for a given positive definite kernel k, the SVM chooses a decision function of the form $f(x) = \sum_i \alpha_i k(x_i, x) + b$ which has a low empirical error $\mathrm{R}_{\mathrm{emp}}$ and is as smooth as possible. According to the large margin point of view, SVMs construct a linear decision boundary in a Hilbert space $\mathcal{H}$ such that the training points are separated with a large margin and the sum of the margin errors is small. Both viewpoints can be connected by embedding the sample space $\mathcal{X}$ into the reproducing kernel Hilbert space $\mathcal{H}$ via the so called "feature map" and the function space $\mathcal{F}$ into the dual $\mathcal{H}'$. Then the regularizer (which is a functional on $\mathcal{F}$) corresponds to the inverse margin (which is a norm of a linear operator), and the empirical error corresponds to the margin error (cf. Sections 4.3 and 7 of Schölkopf and Smola, 2002). The benefits of these two dual viewpoints are that the regularization framework gives some intuition about the geometrical meaning of the norm on $\mathcal{H}$, and the large margin framework leads to statistical learning theory bounds on the generalization error of the classifier.

Now consider the situation where the sample space is a metric space $(\mathcal{X}, d)$. From the regularization point of view, a convenient set of functions on a metric space is the set of Lipschitz functions, as functions with a small Lipschitz constant have low variation. Thus it seems desirable to separate the different classes by a decision function which has a small Lipschitz constant. In this article we want to construct the dual point of view to this approach. To this end, we embed the metric space $(\mathcal{X}, d)$ in a Banach space $\mathcal{B}$ and the space of Lipschitz functions into its dual space $\mathcal{B}'$. Remarkably, both embeddings can be realized as isometries simultaneously. By this construction, each $x \in \mathcal{X}$ will correspond to some $m_x \in \mathcal{B}$ and each Lipschitz function f on $\mathcal{X}$ to some functional $T_f \in \mathcal{B}'$ such that $f(x) = T_f m_x$ and the Lipschitz constant $L(f)$ is equal to the operator norm $\|T_f\|$. In the Banach space $\mathcal{B}$ we can then construct a large margin classifier such that the size of the margin will be given by the inverse of the operator norm of the decision functional. The basic algorithm implementing this approach is

$$\text{minimize } \mathrm{R}_{\mathrm{emp}}(f) + \lambda L(f)$$

in regularization language and

$$\text{minimize } L(f) + C\sum_i \xi_i \text{ subject to } y_i f(x_i) \geq 1 - \xi_i,\ \xi_i \geq 0$$

in large margin language. In both cases, $L(f)$ denotes the Lipschitz constant of the function f, and the minimum is taken over a subset of Lipschitz functions on $\mathcal{X}$. To apply this algorithm in practice, the choice of this subset will be important. We will see that by choosing different subsets we can recover the SVM (in cases where the metric on $\mathcal{X}$ is induced by a kernel), the linear programming machine (cf. Graepel et al., 1999), and even the 1-nearest neighbor classifier. In particular this shows that all these algorithms are large margin algorithms. So the Lipschitz framework can help to analyze a wide range of algorithms which do not seem to be connected at the first glance.

This paper is organized as follows: in Section 2 we provide the necessary functional analytic background for the Lipschitz algorithm, which is then derived in Section 3. We investigate representer theorems for this algorithm in Section 4. It will turn out that the algorithm always has a solution which can be expressed by distance functions to training points. In Section 5 we compute error bounds for the Lipschitz classifier in terms of Rademacher complexities. In particular, this gives valuable information about how fast the algorithm converges for different choices of subsets of Lipschitz functions. The geometrical interpretation for choosing different subsets of Lipschitz functions is further discussed in Section 6.

2. Lipschitz Function Spaces

In this section we introduce several Lipschitz function spaces and their properties. For a comprehensive overview we refer to Weaver (1999).

A metric space $(\mathcal{X}, d)$ is a set $\mathcal{X}$ together with a metric d, that is a non-negative, symmetric function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which fulfills $d(x,y) = 0 \;\Leftrightarrow\; x = y$ and the triangle inequality $d(x,y) +$

$d(y,z) \leq d(x,z)$. A function $f : \mathcal{X} \to \mathbb{R}$ on a metric space $(\mathcal{X},d)$ is called a Lipschitz function if there exists a constant L such that $|f(x) - f(y)| \leq Ld(x,y)$ for all $x,y \in \mathcal{X}$. The smallest constant L such that this inequality holds is called the Lipschitz constant of f, denoted by $L(f)$. For convenience, we recall some standard facts about Lipschitz functions:

Lemma 1 (Lipschitz functions) *Let $(\mathcal{X},d)$ be a metric space, $f,g : \mathcal{X} \to \mathbb{R}$ Lipschitz functions and $a \in \mathbb{R}$. Then $L(f+g) \leq L(f)+L(g)$, $L(af) \leq |a|L(f)$ and $L(\min(f,g)) \leq \max\{L(f),L(g)\}$, where $\min(f,g)$ denotes the pointwise minimum of the functions f and g. Moreover, let $f := \lim_{n\to\infty} f_n$ the pointwise limit of Lipschitz functions f_n with $L(f_n) \leq c$ for all $n \in \mathbb{N}$. Then f is a Lipschitz function with $L(f) \leq c$.*

For a metric space $(\mathcal{X},d)$ consider the set

$$\mathrm{Lip}(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{R};\ f \text{ is a bounded Lipschitz function}\}.$$

It forms a vector space, and the Lipschitz constant $L(f)$ is a seminorm on this space. To define a convenient norm on this space we restrict ourselves to *bounded* metric spaces. These are spaces which have a finite diameter $\mathrm{diam}(\mathcal{X}) := \sup_{x,y\in\mathcal{X}} d(x,y)$. For the learning framework this is not a big drawback as the training and test data can always be assumed to come from a bounded region of the underlying space. For a bounded metric space $\mathcal{X}$ we choose the norm

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

as our default norm on the space $\mathrm{Lip}(\mathcal{X})$. It is easy to see that this indeed is a norm. Note that in the mathematical literature, $\mathrm{Lip}(\mathcal{X})$ is usually endowed with the slightly different norm $\|f\| := \max\{L(f), \|f\|_\infty\}$. But we will see that the norm $\|\cdot\|_L$ fits very naturally in our classification setting, as already can be seen by the following intuitive argument. Functions that are used as classifiers are supposed to take positive and negative values on the respective classes and satisfy

$$\|f\|_\infty = \sup_x |f(x)| \leq \sup_{x,y} |f(x) - f(y)| \leq \mathrm{diam}(\mathcal{X})L(f), \tag{1}$$

that is $\|f\|_L = L(f)$. Hence, the L-norm of a classification decision function is determined by the quantity $L(f)$ we use as regularizer later on. Some more technical reasons for the choice of $\|\cdot\|_L$ will become clear later.

Another important space of Lipschitz functions is constructed as follows. Let $(\mathcal{X}_0,d)$ be a metric space with a distinguished "base point" e which is fixed in advance. $(\mathcal{X}_0,d,e)$ is called a *pointed metric space*. We define

$$\mathrm{Lip}_0(\mathcal{X}_0) := \{f \in \mathrm{Lip}(\mathcal{X}_0);\ f(e) = 0\}.$$

On this space, the Lipschitz constant $L(\cdot)$ is a norm. However, its disadvantage in the learning framework is the condition $f(e) = 0$, which is an inconvenient a priori restriction on our classifier as e has to be chosen in advance. To overcome this restriction, for a given bounded metric space $(\mathcal{X},d)$ we define a corresponding extended pointed metric space $\mathcal{X}_0 := \mathcal{X} \cup \{e\}$ for a new base element e with the metric

$$d_{\mathcal{X}_0}(x,y) = \begin{cases} d(x,y) & \text{for } x,y \in \mathcal{X} \\ \mathrm{diam}(\mathcal{X}) & \text{for } x \in \mathcal{X}, y = e. \end{cases} \tag{2}$$

Note that $\text{diam}(\mathcal{X}_0) = \text{diam}(\mathcal{X})$. Then we define the map

$$\psi : \text{Lip}(\mathcal{X}) \to \text{Lip}_0(\mathcal{X}_0), \quad \psi(f)(x) = \begin{cases} f(x) & \text{if } x \in \mathcal{X} \\ 0 & \text{if } x = e. \end{cases} \tag{3}$$

Lemma 2 (Isometry between Lipschitz function spaces) *ψ is an isometric isomorphism between* $\text{Lip}(\mathcal{X})$ *and* $\text{Lip}_0(\mathcal{X}_0)$.

Proof Obviously, ψ is bijective and linear. Moreover, for $f_0 := \psi(f)$ we have

$$\begin{aligned} L(f_0) &= \sup_{x,y \in \mathcal{X}_0} \frac{|f_0(x) - f_0(y)|}{d_{\mathcal{X}_0}(x,y)} = \max\{\sup_{x,y \in \mathcal{X}} \frac{|f(x) - f(y)|}{d(x,y)}, \sup_{x \in \mathcal{X}} \frac{|f(x) - f(e)|}{d_{\mathcal{X}_0}(x,e)}\} = \\ &= \max\{L(f), \frac{\|f\|_\infty}{\text{diam}(\mathcal{X})}\} = \|f\|_L. \end{aligned}$$

Hence, ψ is an isometry. ■

In some respects, the space $(\text{Lip}_0(\mathcal{X}_0), L(\cdot))$ is more convenient to work with than $(\text{Lip}(\mathcal{X}), \|\cdot\|_L)$. In particular it has some very useful duality properties. Let $(\mathcal{X}_0, d, e)$ be a pointed metric space with some distinguished base element e. A *molecule* of $\mathcal{X}_0$ is a function $m : \mathcal{X}_0 \to \mathbb{R}$ such that its support (i.e., the set where m has non-zero values) is a finite set and $\sum_{x \in \mathcal{X}_0} m(x) = 0$. For $x, y \in \mathcal{X}_0$ we define the *basic molecules* $m_{xy} := \mathbb{1}_x - \mathbb{1}_y$. It is easy to see that every molecule m can be written as a (non unique) finite linear combination of basic molecules. Thus we can define

$$\|m\|_{AE} := \inf \left\{ \sum_i |a_i| d(x_i, y_i);\ m = \sum_i a_i m_{x_i y_i} \right\}$$

which is a norm on the space of molecules. The completion of the space of molecules with respect to $\|\cdot\|_{AE}$ is called the Arens-Eells space $AE(\mathcal{X}_0)$. Denoting its dual space (i.e., the space of all continuous linear forms on $AE(\mathcal{X}_0)$) by $AE(\mathcal{X}_0)'$ the following theorem holds true (cf. Arens and Eells, 1956; Weaver, 1999).

Theorem 3 (Isometry between $AE(\mathcal{X}_0)'$ and $\text{Lip}_0(\mathcal{X}_0)$) *$AE(\mathcal{X}_0)'$ is isometrically isomorphic to* $\text{Lip}_0(\mathcal{X}_0)$.

This means that we can regard a Lipschitz function f on $\mathcal{X}_0$ as a linear functional T_f on the space of molecules, and the Lipschitz constant $L(f)$ coincides with the operator norm of the corresponding functional T_f. For a molecule m and a Lipschitz function f this duality can be expressed as

$$\langle f, m \rangle = \sum_{x \in \mathcal{X}_0} m(x) f(x). \tag{4}$$

It can be proved that $\|m_{xy}\|_{AE} = d(x,y)$ holds for all basic molecules m_{xy}. Hence, it is possible to embed $\mathcal{X}_0$ isometrically in $AE(\mathcal{X}_0)$ via

$$\Gamma : \mathcal{X}_0 \to AE(\mathcal{X}_0),\ x \mapsto m_{xe}. \tag{5}$$

The norm $\|\cdot\|_{AE}$ has a nice geometrical interpretation in terms of the *mass transportation problem* (cf. Weaver, 1999): some product is manufactured in varying amounts at several factories and has to be distributed to several shops. The (discrete) transportation problem is to find an optimal way to transport the product from the factories to the shops. The costs of such a transport are defined as $\sum_{ij} a_{ij}d_{ij}$ where a_{ij} denotes the amount of the product transported from factory i to shop j and d_{ij} the distance between them. If f_i denotes the amount produced in factory i and s_i denotes the amount needed in shop i, the formal definition of the transportation problem is

$$\min_{i,j=1,\ldots,n} \sum a_{ij}d_{ij} \quad \text{subject to} \quad a_{ij} \geq 0, \ \sum_j a_{ij} = s_j, \ \sum_i a_{ij} = f_i. \tag{6}$$

To connect the Arens-Eells space to this problem we identify the locations of the factories and shops with a molecule m. The points x with $m(x) > 0$ represent the factories, the ones with $m(x) < 0$ the shops. It can be proved that $\|m\|_{AE}$ equals the minimal transportation costs for molecule m. A special case is when the given molecule has the form $m_0 = \sum m_{x_i y_j}$. In this case, the transportation problem reduces to the *bipartite minimal matching problem*: given 2m points $(x_1,\ldots,x_n,y_1,\ldots,y_n)$ in a metric space, we want to match each of the x-points to one of the y-points such that the sum of the distances between the matched pairs is minimal. The formal statement of this problem is

$$\min_{\pi} \sum_{i,j} d(x_i, y_{\pi(i)}) \tag{7}$$

where the minimum is taken over all permutations π of the set $\{1,...,n\}$ (cf. Steele, 1997).

In Section 4 we will also need the notion of a vector lattice. A vector lattice is a vector space V with an ordering $\preceq$ which respects the vector space structure (i.e., for $x,y,z \in V, a > 0$: $x \preceq y \implies x + z \preceq y + z$ and $ax \preceq ay$) and such that for any two elements $f,g \in V$ there exists a greatest lower bound $\inf(f,g)$. In particular, the space of Lipschitz functions with the ordering $f \preceq g \Leftrightarrow \forall x \ f(x) \leq g(x)$ forms a vector lattice.

3. The Lipschitz Classifier

Let $(\mathcal{X},d)$ be a metric space and $(x_i,y_i)_{i=1,\ldots,n} \subset \mathcal{X} \times \{\pm 1\}$ some training data. In order to be able to define hyperplanes, we want to embed $(\mathcal{X},d)$ into a vector space, but without loosing or changing the underlying metric structure.

3.1 Embedding and Large Margin in Banach Spaces

Our first step is to embed $\mathcal{X}$ by the identity mapping into the extended space $\mathcal{X}_0$ as described in (2), which in turn is embedded into $AE(\mathcal{X}_0)$ via (5). We denote the resulting composite embedding by

$$\Phi : \mathcal{X} \to AE(\mathcal{X}_0), \ x \mapsto m_x := m_{xe}.$$

Secondly, we identify $\mathrm{Lip}(\mathcal{X})$ with $\mathrm{Lip}_0(\mathcal{X}_0)$ according to (3) and then $\mathrm{Lip}_0(\mathcal{X}_0)$ with $AE(\mathcal{X}_0)'$ according to Theorem 3. Together this defines the map

$$\Psi : \mathrm{Lip}(\mathcal{X}) \to AE(\mathcal{X}_0)', \ f \mapsto T_f.$$

Lemma 4 (Properties of the embeddings) *The mappings Φ and Ψ have the following properties:*

1. *Φ is an isometric embedding of $\mathcal{X}$ into $AE(\mathcal{X}_0)$: to every point $x \in \mathcal{X}$ corresponds a molecule $m_x \in AE(\mathcal{X}_0)$ such that $d(x,y) = \|m_x - m_y\|_{AE}$ for all $x,y \in \mathcal{X}$.*

2. $\mathrm{Lip}(\mathcal{X})$ *is isometrically isomorphic to $AE(\mathcal{X}_0)'$: to every Lipschitz function f on $\mathcal{X}$ corresponds an operator T_f on $AE(\mathcal{X}_0)$ such that $\|f\|_L = \|T_f\|$ and vice versa.*

3. *It makes no difference whether we evaluate operators on the image of $\mathcal{X}$ in $AE(\mathcal{X}_0)$ or apply Lipschitz functions on $\mathcal{X}$ directly: $T_f m_x = f(x)$.*

4. *Scaling a linear operator is the same as scaling the corresponding Lipschitz function: for $a \in \mathbb{R}$ we have $aT_f = T_{af}$.*

Proof All these properties are direct consequences of the construction and Equation (4). ■

The message of this lemma is that it makes no difference whether we classify our training data on the space $\mathcal{X}$ with the decision function $\operatorname{sgn} f(x)$ or on $AE(\mathcal{X}_0)$ with the hyperplane $\operatorname{sgn}(T_f m_x)$. The advantage of the latter is that constructing a large margin classifier in a Banach space is a well studied problem. In Bennett and Bredensteiner (2000) and Zhou et al. (2002) it has been established that constructing a maximal margin hyperplane between the set X^+ of positive and X^- of negative training points in a Banach space V is equivalent to finding the distance between the convex hulls of X^+ and X^-. More precisely, let C^+ and C^- the convex hulls of the sets X^+ and X^-. In the separable case, we define the margin of a separating hyperplane H between C^+ and C^- as the minimal distance between the training points and the hyperplane:

$$\rho(H) := \inf_{i=1,\dots,n} d(x_i, H).$$

The margin of the maximal margin hyperplane coincides with half the distance between the convex hulls of the positive and negative training points. Hence, determining the maximum margin hyperplane can be understood as solving the optimization problem

$$\inf_{p^+ \in C^+, p^- \in C^-} \|p^+ - p^-\|.$$

By duality arguments (cf. Bennett and Bredensteiner, 2000) it can be seen that its solution coincides with the solution of

$$\sup_{T \in V'} \inf_{p^+ \in C^+, p^- \in C^-} \langle T, p^+ - p^- \rangle / \|T\|.$$

This can be equivalently rewritten as the optimization problem

$$\inf_{T \in V', b \in .} \|T\| \text{ subject to } y_i(\langle T, x_i \rangle + b) \geq 1 \;\; \forall i = 1, \dots, n. \tag{8}$$

A solution of this problem is called a large margin classifier. The decision function has the form $f(x) = \langle T, x \rangle + b$, and its margin is given by $1/\|T\|$. For details we refer to Bennett and Bredensteiner (2000) and Zhou et al. (2002).

3.2 Derivation of the Algorithm

Now we can apply this construction to our situation. We embed $\mathcal{X}$ isometrically into the Banach space $AE(\mathcal{X}_0)$ and use the above reasoning to construct a large margin classifier. As the dual space of $AE(\mathcal{X}_0)$ is $\mathrm{Lip}_0(\mathcal{X}_0)$ and $\langle f, m_x \rangle = f(x)$, the optimization problem (8) in our case is

$$\inf_{f_0 \in \mathrm{Lip}_0(\mathcal{X}_0), b \in .} L(f_0) \text{ subject to } y_i(f_0(x_i) + b) \geq 1 \;\; \forall i = 1, ..., n.$$

By the isometry stated in Theorem 3, this is equivalent to the problem

$$\inf_{f \in \mathrm{Lip}(\mathcal{X}), b \in .} \|f\|_L \text{ subject to } y_i(f(x_i) + b) \geq 1 \;\; \forall i = 1, ..., n.$$

Next we want to show that the solution of this optimization problem does not depend on the variable b. To this end, we first set $g := f + b \in \mathrm{Lip}(\mathcal{X})$ to obtain

$$\inf_{g \in \mathrm{Lip}(\mathcal{X}), b \in .} \|g - b\|_L \text{ subject to } y_i g(x_i) \geq 1 \;\; \forall i = 1, ..., n.$$

Then we observe that

$$\|g-b\|_L = \max\{L(g-b), \frac{\|g-b\|_\infty}{\mathrm{diam}(\mathcal{X})}\} = \max\{L(g), \frac{\|g-b\|_\infty}{\mathrm{diam}(\mathcal{X})}\} \geq L(g) = \max\{L(g), \frac{\|g\|_\infty}{\mathrm{diam}(\mathcal{X})}\}.$$

Here the last step is true because of the fact that g takes positive and negative values and thus $\|g\|_\infty / \mathrm{diam}(\mathcal{X}) \leq L(g)$ as we explained in Equation (1) of Section 2. Hence, under the constraints $y_i g(x_i) \geq 1$ we have $\inf_b \|g - b\|_L = L(g)$, and we can rewrite our optimization problem in the final form

$$\inf_{f \in \mathrm{Lip}(\mathcal{X})} L(f) \text{ subject to } y_i f(x_i) \geq 1, \; i = 1, \ldots, n. \qquad (*)$$

We call a solution of this problem a (hard margin) *Lipschitz classifier*. So we have proved:

Theorem 5 (Lipschitz classifier) *Let $(\mathcal{X}, d)$ be a bounded metric space, $(x_i, y_i)_{i=1,...,n} \subset \mathcal{X} \times \{\pm 1\}$ some training data containing points of both classes. Then a solution f of $(*)$ is a large margin classifier, and its margin is given by $1/L(f)$.*

One nice aspect about the above construction is that the margin constructed in the space $AE(\mathcal{X}_0)$ also has a geometrical meaning in the original input space $\mathcal{X}$ itself: it is a lower bound on the minimal distance between the "separation surface" $S := \{s \in \mathcal{X};\; f(s) = 0\}$ and the training points. To see this, normalize the function f such that $\min_{i=1,...,n} |f(x_i)| = 1$. This does not change the set S. Because of

$$1 \leq |f(x_i)| = |f(x_i) - f(s)| \leq L(f) d(x_i, s)$$

we thus get $d(x_i, s) \geq 1/L(f)$.

Analogously to SVMs we also define the soft margin version of the Lipschitz classifier by introducing slack variables ξ_i to allow some training points to lie inside the margin or even be misclassified:

$$\inf_{f \in \mathrm{Lip}(\mathcal{X})} L(f) + C \sum_{i=1}^{n} \xi_i \text{ subject to } y_i f(x_i) \geq 1 - \xi_i, \; \xi_i \geq 0. \qquad (**)$$

In regularization language, the soft margin Lipschitz classifier can be stated as

$$\inf_{f \in \mathrm{Lip}(\mathcal{X})} \ell(y_i f(x_i)) + \lambda L(f)$$

where the loss function ℓ is given by $\ell(y_i f(x_i)) = \max\{0, 1 - y_i f(x_i)\}$.

In Section 4, we will give an analytic expression for a solution of $(*)$ and show how $(**)$ can be written as a linear programming problem. However, it may be sensible to restrict the set over which the infimum is taken in order to avoid overfitting. We thus suggest to consider the above optimization problems over subspaces of $\mathrm{Lip}(\mathcal{X})$ rather than the whole space $\mathrm{Lip}(\mathcal{X})$. In Section 6 we derive a geometrical interpretation of the choice of different subspaces. Now we want to point out some special cases.

Assume that we are given training points in some reproducing kernel Hilbert space H. As it is always the case for linear functions, the Lipschitz constant of a linear function in H' coincides with its Hilbert space norm. This means that the support vector machine in H chooses the same linear function as the Lipschitz algorithm, if the latter takes the subspace of linear functions as hypothesis space.

In the case where we optimize over the subset of all linear combinations of distance functions of the form $f(x) = \sum_{i=1}^n a_i d(x_i, x) + b$, the Lipschitz algorithm can be approximated by the linear programming machine (cf. Graepel et al., 1999):

$$\inf_{a,b} \sum_{i=1}^{n} |a_i| \text{ subject to } y_i(\sum_{i=1}^{n} a_i d(x_i, x) + b) \geq 1.$$

The reason for this is that the Lipschitz constant of a function $f(x) = \sum_{i=1}^n a_i d(x_i, x) + b$ is upper bounded by $\sum_i |a_i|$. Furthermore, if we do not restrict the function space at all, then we will see in the next section that the 1-nearest neighbor classifier is a solution of the Lipschitz algorithm.

These examples show that the Lipschitz algorithm is a very general approach. By choosing different subsets of Lipschitz functions we recover several well known algorithms. As the Lipschitz algorithm is a large margin algorithm according to Theorem 5, the same holds for the recovered algorithms. For instance the linear programming machine, originally designed with little theoretical justification, can now be understood as a large margin algorithm.

4. Representer Theorems

A crucial theorem in the context of SVMs and other kernel algorithms is the representer theorem (cf. Schölkopf and Smola, 2002). It states that even though the space of possible solutions of these algorithms forms an infinite dimensional space, there always exists a solution in the finite dimensional subspace spanned by the training points. It is because of this theorem that SVMs overcome the curse of dimensionality and yield computationally tractable solutions. In this section we prove a similar theorem for the Lipschitz classifiers $(*)$ and $(**)$. To simplify the discussion, we denote $\mathcal{D} := \{d(x, \cdot);\ x \in \mathcal{X}\} \cup \{\mathbb{1}\}$ and $\mathcal{D}_{\text{train}} := \{d(x_i, \cdot);\ x_i \text{ training point }\} \cup \{\mathbb{1}\}$, where $\mathbb{1}$ is the constant-1 function.

4.1 Soft Margin Case

We first start by recalling a general result which implies the classical representer theorem in the case of SVMs.

Lemma 6 (Minimum norm interpolation) *Let V be a function of $n+1$ variables which is non-decreasing in its $n+1$-st argument. Given n points $x_1,\dots,x_n$ and a functional Ω, any function which is a solution of the problem*

$$\inf_f V(f(x_1),\dots,f(x_n),\Omega(f)) \tag{9}$$

is a solution of the minimum norm interpolation problem

$$\inf_{f:\forall i,\, f(x_i)=a_i} \Omega(f) \tag{10}$$

for some $a_1,\dots,a_n \in \mathbb{R}$.

Here, f being a solution of a problem of the form $\inf W(f)$ means $f = \operatorname{argmin} W(f)$. We learned this theorem from M. Pontil, but it seems to be due to C. Micchelli.

Proof Let f_0 be a solution of the first problem. Take $a_i = f_0(x_i)$. Then for any function f such that $f(x_i) = a_i$ for all i, we have

$$V(f(x_1),\dots,f(x_n),\Omega(f)) \geq V(f_0(x_1),\dots,f_0(x_n),\Omega(f_0)) = V(f(x_1),\dots,f(x_n),\Omega(f_0)).$$

Hence, by monotonicity of V we get $\Omega(f) \geq \Omega(f_0)$, which concludes the proof. ■

The meaning of the above result is that if the solutions of problem (10) have specific properties, then the solutions of problem (9) will also have these properties. So instead of studying the properties of solutions of $(**)$ directly, we will investigate the properties of (10) when the functional Ω is the Lipschitz norm. We first need to introduce the concept of Lipschitz extensions.

Lemma 7 (Lipschitz extension) *Given a function f defined on a finite subset $x_1,\dots,x_n$ of $\mathcal{X}$, there exists a function f' which coincides with f on $x_1,\dots,x_n$, is defined on the whole space $\mathcal{X}$, and has the same Lipschitz constant as f. Additionally, it is possible to explicitly construct f' in the form*

$$f'(x) = \alpha \min_{i=1,\dots,n}(f(x_i)+L(f)d(x,x_i)) + (1-\alpha)\max_{i=1,\dots,n}(f(x_i)-L(f)d(x,x_i)),$$

for any $\alpha \in [0,1]$, with $L(f) = \max_{i,j=1,\dots,n}(f(x_i)-f(x_j))/d(x_i,x_j)$.

Proof Consider the function $g(x) = \min_{i=1,\dots,n}(f(x_i)+L(f)d(x,x_i))$. We have

$$|g(x)-g(y)| \leq \max_{i=1,\dots,n} |f(x_i)+L(f)d(x,x_i) - f(x_i) - L(f)d(y,x_i)| \leq L(f)d(x,y),$$

so that $L(g) \leq L(f)$. Also, by definition $g(x_i) \leq f(x_i) + L(f)d(x_i,x_i) = f(x_i)$. Moreover, if i_0 denotes the index where the minimum is achieved in the definition of $g(x_i)$, i.e. $g(x_i) = f(x_{i_0}) + L(f)d(x_i,x_{i_0})$, then by definition of $L(f)$ we have $g(x_i) \geq f(x_{i_0}) + (f(x_i) - f(x_{i_0})) = f(x_i)$. As a result, for all $i = 1,\dots,n$ we have $g(x_i) = f(x_i)$, which also implies that $L(g) = L(f)$.
Now the same reasoning can be applied to $h(x) = \max_{i=1,\dots,n}(f(x_i)-L(f)d(x,x_i))$. Since $\alpha \in [0,1]$ we have $f'(x_i) = f(x_i)$ for all i. Moreover, $L(\alpha g + (1-\alpha)h) \leq \alpha L(g) + (1-\alpha)L(h) = L(f)$ and thus $L(f') = L(f)$, which concludes the proof. ■

From the above lemma, we obtain an easy way to construct solutions of minimum norm interpolation problems like (10) with Lipschitz norms, as is expressed in the next lemma.

Lemma 8 (Solution of the Lipschitz minimal norm interpolation problem) *Let $a_1, \dots, a_n \in \mathbb{R}^n$, $\alpha \in [0,1]$, $L_0 = \max_{i,j=1,\dots,n}(a_i - a_j)/d(x_i, x_j)$, and*

$$f_\alpha(x) := \alpha \min_{i=1,\dots,n}(a_i + L_0 d(x, x_i)) + (1-\alpha) \max_{i=1,\dots,n}(a_i - L_0 d(x, x_i)) .$$

Then f_α is a solution of the minimal norm interpolation problem (10) *with $\Omega(f) = L(f)$. Moreover, when $\alpha = 1/2$ then f_α is a solution of the minimal norm interpolation problem* (10) *with $\Omega(f) = \|f\|_L$.*

Proof Given that a solution f of (10) has to satisfy $f(x_i) = a_i$, it cannot have $L(f) < L_0$. Moreover, by Lemma 7 f_α satisfies the constraints and has $L(f) = L_0$, hence it is a solution of (10) with $\Omega(f) = L(f)$.

When one takes $\Omega(f) = \|f\|_L$, any solution f of (10) has to have $L(f) \geq L_0$ and $\|f\|_\infty \geq \max_i |a_i|$. The proposed solution f_α with $\alpha = 1/2$ not only satisfies the constraints $f_\alpha(x_i) = a_i$ but also has $L(f) = L_0$ and $\|f\|_\infty = \max_i |a_i|$, which shows that it is a solution of the considered problem.

To prove that $\|f\|_\infty = \max_i |a_i|$, consider $x \in \mathcal{X}$ and denote by i_1 and i_2 the indices where the minimum and the maximum, respectively, are achieved in the definition of $f_\alpha(x)$. Then one has

$$f_{1/2}(x) \leq \frac{1}{2}(a_{i_2} + L_0 d(x, x_{i_2})) + \frac{1}{2}(a_{i_2} - L_0 d(x, x_{i_2})) = a_{i_2} ,$$

and similarly $f_{1/2}(x) \geq a_{i_1}$. ■

Now we can formulate a general representer theorem for the soft margin Lipschitz classifier.

Theorem 9 (Soft margin representer theorem) *There exists a solution of the soft margin Lipschitz classifier* $(**)$ *in the vector lattice spanned by $\mathcal{D}_{\text{train}}$ which is of the form*

$$f(x) = \frac{1}{2}\min(a_i + L_0 d(x, x_i)) + \frac{1}{2}\max(a_i - L_0 d(x, x_i))$$

for some real numbers $a_1, \dots, a_n$ with $L_0 := \max_{i,j}(a_i - a_j)/d(x_i, x_j)$. Moreover one has $\|f\|_L = L(f) = L_0$.

Proof The first claim follows from Lemmas 6 and 8. The second claim follows from the fact that a solution of $(**)$ satisfies $\|f\|_L = L(f)$. ■

Theorem 9 is remarkable as the space $\text{Lip}(\mathcal{X})$ of possible solutions of $(**)$ contains the whole vector lattice spanned by $\mathcal{D}$. The theorem thus states that even though the Lipschitz algorithm searches for solutions in the whole lattice spanned by $\mathcal{D}$ it always manages to come up with a solution in the sublattice spanned by $\mathcal{D}_{\text{train}}$.

4.2 Algorithmic Consequences

As a consequence of the above theorem, we can obtain a tractable algorithm for solving problem $(**)$. First, we determine the coefficients a_i by solving

$$\min_{a_1,\dots,a_n \in .} \sum_{i=1}^{n} \ell(y_i a_i) + \lambda \max_{i,j} \frac{(a_i - a_j)}{d(x_i, x_j)} ,$$

which can be rewritten as a linear programming problem

$$\min_{a_1,\ldots,a_n,\xi_1,\ldots,\xi_n,\rho\in} \sum_{i=1}^{n} \xi_i + \lambda\rho\,,$$

under the constraints $\xi_i \geq 0$, $y_i a_i \geq 1-\xi_i$, $\rho \geq (a_i - a_j)/d(x_i,x_j)$. Once a solution is found, one can simply take the function $f_{1/2}$ defined in Theorem 9 with the coefficients a_i determined by the linear program. Note, however, that in practical applications, the solution found by this procedure might overfit as it optimizes $(**)$ over the whole class $\mathrm{Lip}(\mathcal{X})$.

4.3 Hard Margin Case

The representer theorem for the soft margin case clearly also holds in the hard margin case, so that there will always be a solution of $(*)$ in the vector lattice spanned by $\mathcal{D}_{\text{train}}$. But in the hard margin case, also a different representer theorem is valid. We denote the set of all training points with positive label by X^+, the set of the training points with negative label by X^-, and for two subsets $A,B \subset \mathcal{X}$ we define $d(A,B) := \inf_{a\in A, b\in B} d(a,b)$.

Theorem 10 (Hard margin representer theorem) *Problem $(*)$ always has a solution which is a linear combination of distances to* sets *of training points.*

To prove this theorem we first need a simple lemma.

Lemma 11 (Optimal Lipschitz constant) *The Lipschitz constant L^* of a solution of $(*)$ satisfies $L^* \geq 2/d(X^+,X^-)$.*

Proof For a solution f of $(*)$ we have

$$\begin{aligned} L(f) &= \sup_{x,y\in\mathcal{X}} \frac{|f(x)-f(y)|}{d(x,y)} \geq \max_{i,j=1,\ldots,n} \frac{|f(x_i)-f(x_j)|}{d(x_i,x_j)} \\ &\geq \max_{i,j=1,\ldots,n} \frac{|y_i-y_j|}{d(x_i,x_j)} = \frac{2}{\min_{x_i\in X^+, x_j\in X^-} d(x_i,x_j)} = \frac{2}{d(X^+,X^-)}. \end{aligned}$$

■

Lemma 12 (Solutions of $(*)$) *Let $L^* = 2/d(X^+,X^-)$. For all $\alpha \in [0,1]$, the following functions solve $(*)$:*

$$f_\alpha(x) := \alpha \min_i (y_i + L^* d(x,x_i)) + (1-\alpha)\max_i (y_i - L^* d(x,x_i))$$
$$g(x) := \frac{d(x,X^-) - d(x,X^+)}{d(X^+,X^-)}$$

Proof By Lemma 7, f_α has Lipschitz constant L^* and satisfies $f_\alpha(x_i) = y_i$. Moreover, it is easy to see that $y_i g(x_i) \geq 1$. Using the properties of Lipschitz constants stated in Section 2 and the fact that the function $d(x,\cdot)$ has Lipschitz constant 1 we see that $L(g) \leq L^*$. Thus f_α and g are solutions of $(*)$ by Lemma 11. ■

The functions f_α and g lie in the vector lattice spanned by $\mathcal{D}_{\text{train}}$. As g is a linear combination of distances to sets of training points we have proved Theorem 10.

It is interesting to have a closer look at the functions of Lemma 12. The functions f_0 and f_1 are the smallest and the largest functions, respectively, that solve problem $(*)$ with equality in the constraints: any function f that satisfies $f(x_i) = y_i$ and has Lipschitz constant L^* satisfies $f_0(x) \leq f(x) \leq f_1(x)$. The functions g and $f_{1/2}$ are especially remarkable:

Lemma 13 (1-nearest neighbor classifier) *The functions g and $f_{1/2}$ defined above have the sign of the* 1*-nearest neighbor classifier.*

Proof It is obvious that $g(x) > 0 \iff d(x,X^+) < d(x,X^-)$ and $g(x) < 0 \iff d(x,X^+) > d(x,X^-)$. For the second function, we rewrite $f_{1/2}$ as follows:

$$f_{1/2}(x) = \frac{1}{2}(\min(L^*d(x,X^+)+1, L^*d(x,X^-)-1) - \min(L^*d(x,X^+)-1, L^*d(x,X^-)+1)) .$$

Consider x such that $d(x,X^+) \geq d(x,X^-)$. Then $d(x,X^+)+1 \geq d(x,X^-)-1$ and thus

$$f_{1/2}(x) = \frac{1}{2}\left(L^*d(x,X^-) - 1 - \min(L^*d(x,X^+)-1, L^*d(x,X^-)+1)\right) \leq 0 .$$

The same reasoning applies to the situation $d(x,X^+) \leq d(x,X^-)$ to yield $f_{1/2}(x) \geq 0$ in this case. ■

Note that g needs not reach equality in the constraints on all the data points, whereas the function $f_{1/2}$ always satisfies equality in the constraints. Lemma 13 has the surprising consequence that according to Section 3, the 1-nearest neighbor classifier actually is a large margin classifier.

4.4 Negative Results

So far we have proved that $(*)$ always has a solution which can be expressed as a linear combination of distances to sets of training points. But maybe we even get a theorem stating that we always find a solution which is a linear combination of distance functions to single training points? Unfortunately, in the metric space setting such a theorem is not true in general. This can be seen by the following counterexample:

Example 1 *Assume four training points x_1, x_2, x_3, x_4 with distance matrix*

$$D = \begin{pmatrix} 0 & 2 & 1 & 1 \\ 2 & 0 & 1 & 1 \\ 1 & 1 & 0 & 2 \\ 1 & 1 & 2 & 0 \end{pmatrix}$$

and label vector $y = (1,1,-1,-1)$. Then the set

$$\{f : \mathcal{X} \to \mathbb{R} |\, y_i f(x_i) \geq 1,\ f(x) = \sum_{i=1}^{4} a_i d(x_i,x) + b\}$$

is empty. The reason for this is that the distance matrix is singular and we have $d(x_1,\cdot)+d(x_2,\cdot) = d(x_3,\cdot) = d(x_4,\cdot)$. *Hence, in this example,* $(*)$ *has no solution which is a linear combination of distances to single training points. But it still has a solution as linear combination of distances to sets of training points according to Theorem 10.*

Another negative result is the following. Assume that instead of looking for solutions of $(*)$ in the space of all Lipschitz functions we only consider functions in the vector space spanned by $\mathcal{D}$. Is it in this case always possible to find solution in the linear span of $\mathcal{D}_{train}$? The answer is no again. An example for this is the following:

Example 2 *Let* $\mathcal{X} = \{x_1,...,x_5\}$ *consist of five points with distance matrix*

$$D = \begin{pmatrix} 0 & 2 & 1 & 1 & 1 \\ 2 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 \\ 1 & 1 & 2 & 0 & 2 \\ 1 & 1 & 1 & 2 & 0 \end{pmatrix}.$$

Let the first four points be training points with the label vector $y = (-1,-1,-1,1)$. *As above there exists no feasible function in the vector space spanned by* $\mathcal{D}_{train}$. *But as the distance matrix of all five points is invertible, there exist feasible functions in the vector space spanned by* $\mathcal{D}$.

In the above examples the problem was that the distance matrix on the training points was singular. But there are also other sources of problems that can occur. In particular it can be the case that the Lipschitz constant of a function restricted to the training set takes the minimal value L^*, but the Lipschitz constant on the whole space $\mathcal{X}$ is larger. Then it can happen that although we can find a linear combination of distance functions that satisfies $f(x_i) = y_i$, the function f has a Lipschitz constant larger than L^* and thus is no solution of $(*)$. An example for this situation is the following:

Example 3 *Let* $\mathcal{X} = \{x_1,...,x_5\}$ *consist of five points with distance matrix*

$$D = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 2 \\ 1 & 1 & 0 & 2 & 1 \\ 1 & 1 & 2 & 0 & 1 \\ 1 & 2 & 1 & 1 & 0 \end{pmatrix}.$$

Let the first four points be training points with the label vector $y = (1,1,-1,-1)$. *The optimal Lipschitz constant in this problem is* $L^* = 2/d(X^+,X^-) = 2$. *The function* $f(x) = -2d(x_1,x) - 2d(x_2,x) + 3$ *has this Lipschitz constant if we evaluate it on the training points only. But if we also consider* x_5, *the function has Lipschitz constant* 4.

These examples show that, in general, Theorem 10 cannot be improved to work in the vector space instead of the vector lattice spanned by $\mathcal{D}_{\text{train}}$. This also holds if we consider some subspaces of the set of Lipschitz functions. Thus we are in the interesting situation that it is not enough to consider distance functions to single training points – we have to deal with distances to sets of training points.

5. Error Bounds

In this section we compute error bounds for the Lipschitz classifier using Rademacher averages. This can be done following techniques introduced for example in Chapter 3 of Devroye and Lugosi (2001) or in Bartlett and Mendelson (2002). The measures of capacity we consider are the Rademacher average R_n and the related maximum discrepancy $\tilde{R}_n$. For an arbitrary class $\mathcal{F}$ of functions, they are defined as

$$R_n(\mathcal{F}) := E\left(\frac{1}{n}\sup_{f\in\mathcal{F}}|\sum_{i=1}^{n}\sigma_i f(X_i)|\right) \geq \frac{1}{2}E\left(\frac{1}{n}\sup_{f\in\mathcal{F}}|\sum_{i=1}^{n}(f(X_i)-f(X_i'))\right)| =: \frac{1}{2}\tilde{R}_n(\mathcal{F})$$

where σ_i are iid Rademacher random variables (i.e., $Prob(\sigma_i=+1)=Prob(\sigma_i=-1)=1/2$), X_i and X_i' are iid sample points according to the (unknown) sample distribution, and the expectation is taken with respect to all occurring random variables. Sometimes we also consider the conditional Rademacher average $\hat{R}_n$, where the expectation is taken only conditionally on the sample points $X_1,\ldots,X_n$. For decision function f, consider the loss function $\ell(f(x),y)=1$ if $yf(x)\leq -1$, $1-yf(x)$ if $0\leq yf(x)\leq 1$, and 0 if $yf(x)\geq 1$. Let $\mathcal{F}$ be a class of functions, denote by E the expectation with respect to the unknown sample distribution and by E_n the expectation with respect to the empirical distribution of the training points.

Lemma 14 (Error bounds) *With probability at least* $1-\delta$ *over the iid drawing of n sample points, every* $f\in\mathcal{F}$ *satisfies*

$$E(\ell(f(X),Y)) \leq E_n(\ell(f(X),Y)) + 2R_n(\mathcal{F}) + \sqrt{\frac{8\log(2/\delta)}{n}}.$$

Proof The proof is based on techniques of Devroye and Lugosi (chap. 3 of 2001) and Bartlett and Mendelson (2002): McDiarmid's concentration inequality, symmetrization and contraction property of Rademacher averages. ■

A similar bound can be obtained with the maximum discrepancy (see Bartlett and Mendelson, 2002).

We will describe two different ways to compute Rademacher averages for sets of Lipschitz functions. One way is a classical approach using entropy numbers and leads to an upper bound on R_n. For this approach we always assume that the metric space $(\mathcal{X},d)$ is precompact (i.e., it can be covered by finitely many balls of radius ε for every $\varepsilon>0$).

The other way is more elegant: because of the definition of $\|\cdot\|_L$ and the resulting isometries, the maximum discrepancy of a $\|\cdot\|_L$-unit ball of $\text{Lip}(\mathcal{X})$ is the same as of the corresponding unit ball in $AE(\mathcal{X}_0)'$. Hence it will be possible to express $\tilde{R}_n$ as the norm of an element of the Arens-Eells space. This norm can then be computed via bipartite minimal matching. In the following, B always denotes the unit ball of the considered function space.

5.1 The Duality Approach

The main insight to compute the maximum discrepancy by the duality approach is the following observation:

$$\sup_{\|f\|_L \le 1} |\sum_{i=1}^n f(x_i) - f(x_i')| = \sup_{\|T_f\| \le 1} |\sum_{i=1}^n T_f m_{x_i} - T_f m_{x_i'}| =$$
$$= \sup_{\|T_f\| \le 1} |\langle T_f, \sum_{i=1}^n m_{x_i} - m_{x_i'} \rangle| = \| \sum_{i=1}^n m_{x_i x_i'} \|_{AE}$$

Applying this to the definition of the maximum discrepancy immediately yields

$$\tilde{R}_n(B) = \frac{1}{n} E \| \sum_{i=1}^n m_{X_i X_i'} \|_{AE}. \tag{11}$$

As we already explained in Section 2, the norm $\| \sum_{i=1}^n m_{X_i X_i'} \|_{AE}$ can be interpreted as the costs of a minimal bipartite matching between $\{X_1, \ldots, X_n\}$ and $\{X_1', \ldots, X_n'\}$. To compute the right hand side of (11) we need to know the expected value of random instances of the bipartite minimal matching problem, where we assume that the points X_i and X_i' are drawn iid from the sample distribution. In particular we want to know how this value scales with the number n of points as this indicates how fast we can learn. This question has been solved for some special cases of random bipartite matching. Let the random variable C_n describe the minimal bipartite matching costs for a matching between the points $X_1, \ldots, X_n$ and $X_1', \ldots, X_n'$ drawn iid according to some distribution P. In Dobric and Yukich (1995) it has been proved that for an arbitrary distribution on the unit square of $\mathbb{R}^d$ with $d \ge 3$ we have $\lim C_n/(n^{d-1/d}) = c > 0$ a.s. for some constant c. The upper bound $EC_n \le c\sqrt{n \log n}$ for arbitrary distributions on the unit square in $\mathbb{R}^2$ was presented in Talagrand (1992). These results, together with Equation (11), lead to the following maximum discrepancies:

Theorem 15 (Maximum discrepancy of unit ball of $\mathrm{Lip}([0,1]^d)$**)** *Let $\mathcal{X} = [0,1]^d \subset \mathbb{R}^d$ with the Euclidean metric. Then the maximum discrepancy of the $\| \cdot \|_L$-unit ball B of $\mathrm{Lip}(\mathcal{X})$ satisfies*

$$\tilde{R}_n(B) \le c_2 \sqrt{\log n}/\sqrt{n} \quad \textit{for all } n \in \mathbb{N} \qquad \textit{if } d = 2$$
$$\lim_{n \to \infty} \tilde{R}_n(B) \sqrt[d]{n} = c_d > 0 \qquad \textit{if } d \ge 3$$

where c_d $(d \ge 2)$ are constants which are independent of n but depend on d.

Note that this procedure gives (asymptotically) exact results rather than upper bounds in cases where we have (asymptotically) exact results on the bipartite matching costs. This is for example the case for cubes in $\mathbb{R}^d, d \ge 3$ as Dobric and Yukich (1995) gives an exact limit result, or for $\mathbb{R}^2$ with the uniform distribution.

5.2 Covering Number Approach

To derive the Rademacher complexity in more general settings than Euclidean spaces we use an adapted version of the classical entropy bound of Dudley based on covering numbers. The covering number $N(\mathcal{X}, \varepsilon, d)$ of a totally bounded metric space $(\mathcal{X}, d)$ is the smallest number of balls of radius ε with centers in $\mathcal{X}$ which can cover $\mathcal{X}$ completely. The proof of the following theorem can be found in the appendix.

Theorem 16 (Generalized entropy bound) *Let $\mathcal{F}$ be a class of functions and $X_1,\dots,X_n$ iid sample points with empirical distribution μ_n. Then, for every $\varepsilon > 0$,*

$$\hat{R}_n(\mathcal{F}) \leq 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\varepsilon/4}^{\infty} \sqrt{\log N(\mathcal{F}, u, L_2(\mu_n))}\, du.$$

To apply this theorem we need to know covering numbers of spaces of Lipschitz functions. This can be found for example in Kolmogorov and Tihomirov (1961), pp.353–357.

Theorem 17 (Covering numbers for Lipschitz function balls) *For a totally bounded metric space $(\mathcal{X}, d)$ and the unit ball B of $(\mathrm{Lip}(\mathcal{X}), \|\cdot\|_L)$,*

$$2^{N(\mathcal{X}, 4\varepsilon, d)} \leq N(B, \varepsilon, \|\cdot\|_\infty) \leq \left(2\left\lceil \frac{2\,\mathrm{diam}(\mathcal{X})}{\varepsilon}\right\rceil + 1\right)^{N(\mathcal{X}, \frac{\varepsilon}{4}, d)}.$$

If, in addition, $\mathcal{X}$ is connected and centered (i.e., for all subsets $A \subset \mathcal{X}$ with $\mathrm{diam}(A) \leq 2r$ there exists a point $x \in \mathcal{X}$ such that $d(x,a) \leq r$ for all $a \in A$),

$$2^{N(\mathcal{X}, 2\varepsilon, d)} \leq N(B, \varepsilon, \|\cdot\|_\infty) \leq \left(2\left\lceil \frac{2\,\mathrm{diam}(\mathcal{X})}{\varepsilon}\right\rceil + 1\right) \cdot 2^{N(\mathcal{X}, \frac{\varepsilon}{2}, d)}.$$

Combining Theorems 16 and 17 and using $N(\mathcal{F}, u, L_2(\mu_n)) \leq N(\mathcal{F}, u, \|\cdot\|_\infty)$ now gives a bound on the Rademacher complexity of balls of $\mathrm{Lip}(\mathcal{X})$:

Theorem 18 (Rademacher complexity of unit ball of $\mathrm{Lip}(\mathcal{X})$) *Let $(\mathcal{X}, d)$ be a totally bounded metric space with diameter $\mathrm{diam}(\mathcal{X})$ and B the ball of Lipschitz functions with $\|f\|_L \leq 1$. Then, for every $\varepsilon > 0$,*

$$R_n(B) \leq 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\varepsilon/4}^{4\,\mathrm{diam}(\mathcal{X})} \sqrt{N(\mathcal{X}, \frac{u}{4}, d) \log\left(2\left\lceil \frac{2\,\mathrm{diam}(\mathcal{X})}{u}\right\rceil + 1\right)}\, du.$$

If, in addition, $\mathcal{X}$ is connected and centered, we have

$$R_n(B) \leq 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\varepsilon/4}^{2\,\mathrm{diam}(\mathcal{X})} \sqrt{N(\mathcal{X}, \frac{u}{2}, d) \log 2 + \log(2\left\lceil \frac{2\,\mathrm{diam}(\mathcal{X})}{u}\right\rceil + 1)}\, du.$$

In our framework this is a nice result as the bound on the complexity of balls of $\mathrm{Lip}(\mathcal{X})$ only uses the metric properties of the underlying space $\mathcal{X}$. Now we want to compare the results of Theorems 15 and 18 for two simple examples.

Example 4 (d-dimensional unit square, $d \geq 3$) *Let $\mathcal{X} = [0,1]^d \subset \mathbb{R}^d, d \geq 3$, with the Euclidean metric $\|\cdot\|_2$. This is a connected and centered space. In Theorem 15 we showed that $\tilde{R}_n(B)$ asymptotically scales as $1/\sqrt[d]{n}$, and this result cannot be improved. Now we want to check whether Theorem 18 achieves a similar scaling rate. To this end we choose $\varepsilon = 1/\sqrt[d]{n}$ (as we know that we cannot obtain a rate smaller than this) and use that the covering numbers of $\mathcal{X}$ have the form $N(\mathcal{X}, \varepsilon, \|\cdot\|_2) = c/\varepsilon^d$ (e.g., page 1 of Mendelson and Vershynin, 2003). After evaluating the second integral of Theorem 18 we find that $R_n(B)$ indeed scales as $1/\sqrt[d]{n}$.*

Example 5 (2-dimensional unit square) *Let $\mathcal{X} = [0,1]^2 \subset \mathbb{R}^2$ with the Euclidean metric. Applying Theorem 18 similar to Example 4 yields a bound on $R_n(B)$ that scales as $\log n/\sqrt{n}$.*

In case of Example 4 the scaling behavior of the upper bound on $R_n(B)$ obtained by the covering number approach coincides with the exact result for $\tilde{R}_n(B)$ derived in Theorem 15. In case of Example 5 the covering number result $\log n/\sqrt{n}$ is slightly worse than the result $\sqrt{\log(n)}/\sqrt{n}$ obtained in Theorem 15.

5.3 Complexity of Lipschitz RBF Classifiers

In this section we want to derive a bound for the Rademacher complexity of radial basis function classifiers of the form

$$\mathcal{F}_{rbf} := \{f : \mathcal{X} \to \mathbb{R} |\ f(x) = \sum_{k=1}^{l} a_k g_k(d(p_k,x)),\ g_k \in \mathcal{G},\ l < \infty\}, \tag{12}$$

where $p_k \in \mathcal{X}$, $a_k \in \mathbb{R}$, and $\mathcal{G} \subset \mathrm{Lip}(\mathcal{X})$ is a (small) set of $\|\cdot\|_\infty$-bounded Lipschitz functions on $\mathbb{R}$ whose Lipschitz constants are bounded from below by a constant $c > 0$. As an example, consider $\mathcal{G} = \{g : \mathbb{R} \to \mathbb{R} |\ g(x) = \exp(-x^2/\sigma^2), \sigma \geq 1\}$. The special case $\mathcal{G} = \{id\}$ corresponds to the function class which is used by the linear programming machine. It can easily be seen that the Lipschitz constant of an RBF function satisfies $L(\sum_k a_k g_k(d(p_k,\cdot))) \leq \sum_k |a_k| L(g_k)$. We define a norm on $\mathcal{F}_{rbf}$ by

$$\|f\|_{rbf} := \inf\left\{\sum_k |a_k| L(g_k);\ f = \sum_k a_k g_k(d(p_k,\cdot))\right\}$$

and derive the Rademacher complexity of a unit ball B of $(\mathcal{F}_{rbf}, \|\cdot\|_{rbf})$. Substituting a_k by $c_k/L(g_k)$ in the expansion of f we get

$$\begin{aligned}
\sup_{f\in B} |\sum_{i=1}^{n} \sigma_i f(x_i)| &= \sup_{\sum |a_k| L(g_k) \leq 1, p_k \in \mathcal{X}, g_k \in \mathcal{G}} |\sum_{i=1}^{n} \sigma_i \sum_{k=1}^{l} a_k g_k(d(p_k,x_i))| \\
&= \sup_{\sum |c_k| \leq 1, p_k \in \mathcal{X}, g_k \in \mathcal{G}} |\sum_{i=1}^{n} \sigma_i \sum_{k=1}^{l} \frac{c_k}{L(g_k)} g_k(d(p_k,x_i))| \\
&= \sup_{\sum |c_k| \leq 1, p_k \in \mathcal{X}, g_k \in \mathcal{G}} |\sum_{k=1}^{l} c_k \sum_{i=1}^{n} \sigma_i \frac{1}{L(g_k)} g_k(d(p_k,x_i))| \\
&= \sup_{p\in\mathcal{X}, g\in\mathcal{G}} |\sum_{i=1}^{n} \sigma_i \frac{1}{L(g)} g(d(p,x_i))|.
\end{aligned} \tag{13}$$

For the last step observe that the supremum in the linear expansion in the second last line is obtained when one of the c_k is 1 and all the others are 0. To proceed we introduce the notations $h_{p,g}(x) := g(d(p,x_i))/L(g)$, $\mathcal{H} := \{h_{p,g};\ p \in \mathcal{X}, g \in \mathcal{G}\}$, and $\mathcal{G}_1 := \{g/L(g);\ g \in \mathcal{G}\}$. We rewrite the right hand side of Equation (13) as

$$\sup_{p\in\mathcal{X}, g\in\mathcal{G}} |\sum_{i=1}^{n} \sigma_i \frac{1}{L(g)} g(d(p,x_i))| = \sup_{h_{p,g}\in\mathcal{H}} |\sum_{i=1}^{n} \sigma_i h_{p,g}(x_i)|$$

and thus obtain $R_n(B) = R_n(\mathcal{H})$. To calculate the latter we need the following:

Lemma 19 $N(\mathcal{H}, 2\varepsilon, \|\cdot\|_\infty) \leq N(\mathcal{X}, \varepsilon, d) N(\mathcal{G}_1, \varepsilon, \|\cdot\|_\infty)$.

Proof First we observe that for $h_{p_1,g_1}, h_{p_2,g_2} \in \mathcal{H}$

$$\begin{aligned} \|h_{p_1,g_1} - h_{p_2,g_2}\|_\infty &= \sup_{x\in\mathcal{X}} |\frac{g_1(d(p_1,x))}{L(g_1)} - \frac{g_2(d(p_2,x))}{L(g_2)}| \\ &\leq \sup_{x\in\mathcal{X}} \left(|\frac{g_1(d(p_1,x))}{L(g_1)} - \frac{g_1(d(p_2,x))}{L(g_1)}| + |\frac{g_1(d(p_2,x))}{L(g_1)} - \frac{g_2(d(p_2,x))}{L(g_2)}| \right) \\ &\leq \sup_{x\in\mathcal{X}} |d(p_1,x) - d(p_2,x)| + \|\frac{g_1}{L(g_1)} - \frac{g_2}{L(g_2)}\|_\infty \\ &\leq d(p_1,p_2) + \|\frac{g_1}{L(g_1)} - \frac{g_2}{L(g_2)}\|_\infty =: d_{\mathcal{H}}(h_{p_1,g_1}, h_{p_2,g_2}) \end{aligned} \qquad (14)$$

For the step from the second to the third line we used the Lipschitz property of g_1. Finally, it is easy to see that $N(\mathcal{H}, 2\varepsilon, d_{\mathcal{H}}) \leq N(\mathcal{X}, \varepsilon, d) N(\mathcal{G}_1, \varepsilon, \|\cdot\|_\infty)$. ■

Plugging lemma 19 in Theorem 16 yields the following Rademacher complexity:

Theorem 20 (Rademacher complexity of unit ball of $\mathcal{F}_{rbf}$) *Let B be the unit ball of $(\mathcal{F}_{rbf}, \|\cdot\|_{rbf})$, $\mathcal{G}_1$ the rescaled functions of $\mathcal{G}$ as defined above, and $w := \max\{\text{diam}(\mathcal{X}, d), \text{diam}(\mathcal{G}_1, \|\cdot\|_\infty)\}$. Then, for every $\varepsilon > 0$,*

$$R_n(B) \leq 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\varepsilon/4}^{w} \sqrt{\log N(\mathcal{X}, \frac{u}{2}, d) + \log N(\mathcal{G}_1, \frac{u}{2}, \|\cdot\|_\infty)}\ du.$$

This theorem is a huge improvement compared to Theorem 18 as instead of the covering numbers we now have log-covering numbers in the integral. As an example consider the linear programming machine on $\mathcal{X} = [0,1]^d$. Because of $\mathcal{G} = \{id\}$, the second term in the square root vanishes, and the integral over the log-covering numbers of $\mathcal{X}$ can be bounded by a constant independent of ε. As result we obtain that in this case $R_n(B)$ scales as $1/\sqrt{n}$.

6. Choosing Subspaces of $\text{Lip}(\mathcal{X})$

So far we always considered the isometric embedding of the given metric space into the Arens-Eells space and discovered many interesting properties of this embedding. But there exist many different isometric embeddings which could be used instead. Hence, the construction of embedding the metric space isometrically into some Banach space and then using a large margin classifier in this Banach space is also possible with different Banach spaces than the Arens-Eells space. For example, Hein and Bousquet (2003) used the Kuratowski embedding, which maps a metric space $\mathcal{X}$ isometrically in the space of continuous functions $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ (see Example 6 below). Now it is a natural question whether there are interesting relationships between large margin classifiers constructed by the different isometric embeddings, especially with respect to the Lipschitz classifier.

A second question concerns the choice of subspaces of $\text{Lip}(\mathcal{X})$. At the end of Section 3 we already explained that we have to work on some "reasonable" subspace of Lipschitz functions to apply the Lipschitz classifier in practice. This is justified by complexity arguments, but does the large margin interpretation still hold if we do this? Is there some geometric intuition which could help choosing a subspace?

It will turn out that both questions are inherently related to each other. We will show that there is a correspondence between embedding $\mathcal{X}$ into a Banach space V and constructing the large margin classifier on V on the one hand, and choosing a subspace F of $\text{Lip}(\mathcal{X})$ and constructing the Lipschitz classifier from F on the other hand. Ideally, we would like to have a one-to-one correspondence between V and F. In one direction this would mean that we could realize any large margin classifier on any Banach space V with the Lipschitz classifier on an appropriate subspace F of Lipschitz functions. In the other direction this would mean that choosing a subspace F of Lipschitz functions corresponds to a large margin classifier on some Banach space V. We could then study the geometrical implications of a certain subspace F via the geometric properties of V.

Unfortunately, such a nice one-to-one correspondence between V and F is not always true, but in many cases it is. We will show that given an embedding into some vector space V, the hypothesis class of the large margin classifier on V always corresponds to a subspace F of Lipschitz functions (Lemma 24). In general, this correspondence will be an isomorphism, but not an isometry. The other way round, given a subspace F of Lipschitz functions, under some conditions we can construct a vector space V such that $\mathcal{X}$ can be isometrically embedded into V and the large margin classifiers on V and F coincide (Lemma 25).

The key ingredient in this section is the fact that $AE(\mathcal{X}_0)$ is a free Banach space. The following definition can be found for example in Pestov (1986).

Definition 21 (Free Banach space) *Let $(\mathcal{X}_0, d, e)$ be a pointed metric space. A Banach space $(E, \|\cdot\|_E)$ is a free Banach space over $(\mathcal{X}_0, d, e)$ if the following properties hold:*

1. *There exists an isometric embedding $\Phi : \mathcal{X}_0 \to E$ with $\Phi(e) = 0$, and E is the closed linear span of $\Phi(\mathcal{X}_0)$.*
2. *For every Banach space $(V, \|\cdot\|_V)$ and every Lipschitz map $\Psi : \mathcal{X}_0 \to V$ with $L(\Psi) = 1$ and $\Psi(e) = 0$ there exists a linear operator $T : E \to V$ with $\|T\| = 1$ such that $T \circ \Phi = \Psi$.*

It can be shown that the free Banach space over $(\mathcal{X}, d, e)$ always exists and is unique up to isomorphism (cf. Pestov, 1986).

Lemma 22 (*AE* is a free Banach space) *For any pointed metric space $(\mathcal{X}_0, d, e)$, $AE(\mathcal{X}_0)$ is a free Banach space.*

Proof Property (1) of Definition 21 is clear by construction. For a proof of property (2), see for example Theorem 2.2.4 of Weaver (1999). ■

We are particularly interested in the case where the mapping $\Psi : \mathcal{X}_0 \to V$ of Definition 21 is an isometric embedding of $\mathcal{X}_0$ into some vector space V. Firstly we want to find out under which conditions its dual V' is isometric isomorphic to some subspace F of $\text{Lip}(\mathcal{X})$. Secondly, given a subspace F of $\text{Lip}(\mathcal{X})$ the question is whether there exists a Banach space V such that $\mathcal{X}_0$ can be embedded isometrically into V and simultaneously V' is isometric isomorphic to F. Both questions will be answered by considering the mapping T of Definition 21 and its adjoint T'. The following treatment will be rather technical, and it might be helpful to have Figure 1 in mind, which shows which relations we want to prove.

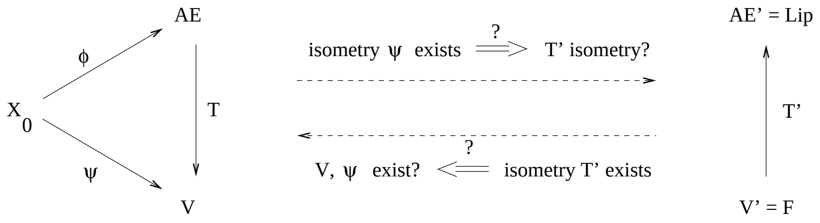
AE
φ
X0
ψ
V
T
isometry ψ exists ?⟹ T' isometry?
V, ψ exist? ?⟸ isometry T' exists
AE' = Lip
T'
V' = F

Figure 1: *Relations between Banach spaces and subspaces of Lipschitz functions. The left part shows the commutative diagram corresponding to the free Banach space property of $AE(\mathcal{X}_0)$. The right part shows the adjoint mapping T' of T. The dotted arrows in the middle show the relationships we want to investigate.*

Now we want to go into detail and start with the first question. For simplicity, we make the following definition.

Definition 23 (Dense isometric embedding) *Let $(\mathcal{X}_0, d)$ a metric space and V a normed space. A mapping $\Psi : \mathcal{X}_0 \to V$ is called a* dense isometric embedding *if Ψ is an isometry and if V is the norm-closure of* $\text{span}\{\Psi(x); x \in \mathcal{X}_0\}$.

Lemma 24 (Construction of F for given V) *Let $(\mathcal{X}_0, d)$ be a pointed metric space, $(V, \|\cdot\|_V)$ a normed space and $\Psi : \mathcal{X}_0 \to V$ a dense isometric embedding. Then V' is isomorphic to a closed subspace $F \subset \text{Lip}_0(\mathcal{X}_0)$, and the canonical injection $i : F \to \text{Lip}_0(\mathcal{X}_0)$ satisfies $\|i\| \leq 1$.*

Proof Recall the notation $m_x := \Phi(x)$ from Section 3 and analogously denote $v_x := \Psi(x)$. Let $T : AE(\mathcal{X}_0) \to V$ the linear mapping with $T \circ \Phi = \Psi$ as in Definition 21. As Ψ is an isometry, T satisfies $\|T\| = 1$, and maps $AE(\mathcal{X}_0)$ on some dense subspace of V. Consider the adjoint $T' : V' \to AE(\mathcal{X}_0)'$. It is well known (e.g., Chapter 4 of Rudin, 1991) that $\|T\| = \|T'\|$ and that T' is injective iff the range of T is dense. Thus, in our case T' is injective. As by construction also $\langle Tm_x, v' \rangle = \langle T'v', m_x \rangle$, we have a unique correspondence between the linear functions in V' and some subspace $F := T'V' \subset AE(\mathcal{X}_0)'$: for $g \in V'$ and $f = T'g \in \text{Lip}_0(\mathcal{X}_0)$ we have $g(v_x) = f(m_x)$ for every $x \in \mathcal{X}_0$. The canonical inclusion i corresponds to the adjoint T'. ■

Lemma 24 shows that the hypothesis space V' constructed by embedding $\mathcal{X}$ into V is isomorphic to a subset $F \subset \text{Lip}_0(\mathcal{X}_0)$. But it is important to note that this isomorphism is not isometric in general. Let $g \in V'$ and $f \in \text{Lip}_0(\mathcal{X}_0)$ be corresponding functions, that is $f = T'g$. Because of $\|T'\| = 1$ we know that $\|f\|_{AE'} \leq \|g\|_V$, but in general we do not have equality. This means that the margins $\|g\|_{V'}$ and $\|f\|_{AE'}$ of corresponding functions are measured with respect to different norms and might have different sizes. As a consequence, the solutions of the two large margin problems

$$\min_{g \in V'} \|g\|_{V'} \text{ subject to } y_i g(v_{x_i}) \geq 1$$

and

$$\min_{f \in F} \|f\|_L \text{ subject to } y_i f(x_i) \geq 1$$

might be different, even though the sets of feasible functions are the same in both cases.

To illustrate this we will consider two examples. The first one shows how the large margin classifier in V can give different results than the one constructed by using the corresponding subspace for the Lipschitz classifier. In the second example we show a situation where both classifiers coincide.

Example 6 (Kuratowski embedding) *Let $(\mathcal{X}, d)$ be an arbitrary compact metric space and $(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)$ the space of continuous functions on $\mathcal{X}$. Define $\Psi : \mathcal{X} \to \mathcal{C}(\mathcal{X})$, $x \mapsto d(x, \cdot)$. This mapping is an isometric embedding called Kuratowski embedding, and it has been used in Hein and Bousquet (2003) to construct a large margin classifier. We want to compare the large margin classifiers resulting from the Kuratowski embedding and the embedding in the Arens-Eells space. As an example consider the finite metric space $\mathcal{X} = \{x_1, ..., x_4\}$ with distance matrix*

$$D = \begin{pmatrix} 0 & 5 & 3 & 6 \\ 5 & 0 & 4 & 1 \\ 3 & 4 & 0 & 5 \\ 6 & 1 & 5 & 0 \end{pmatrix}.$$

Let $V = \mathrm{span}\{d(x, \cdot);\ x \in \mathcal{X}\} \subset \mathcal{C}(\mathcal{X})$, endowed with the norm $\|\cdot\|_\infty$. V is a 4-dimensional vector space. Let V' its dual space. Via the mapping T', each linear operator $g \in V'$ corresponds to the linear operator $f \in \mathrm{Lip}_0(\mathcal{X}_0)$ with $f(x_i) = \langle g, d(x_i, \cdot)\rangle =: c_i$. Now we want to compare the norms of g in V' and f in $\mathrm{Lip}(\mathcal{X})$.The norm of g in V' can be computed as follows:

$$\begin{aligned} \|g\|_{V'} &= \sup\{\langle g, v\rangle :\ v \in V,\ \|v\|_V \leq 1\} \\ &= \sup\{\langle g, \sum_{i=1}^4 a_i d(x_i, \cdot)\rangle :\ a_i \in \mathbb{R},\ \|\sum_{i=1}^4 a_i d(x_i, \cdot)\|_\infty \leq 1\} \\ &= \sup\{\sum_{i=1}^4 a_i c_i :\ a_i \in \mathbb{R},\ -1 \leq \sum_{i=1}^4 a_i d(x_i, x_j) \leq 1 \textit{ for all } j = 1, ..., 4\ \}. \end{aligned}$$

For given function $g \in V'$ (that is, for given values c_i) this norm can be computed by a linear program. Consider the two functions $g_1, g_2 \in V'$ with values on x_1, x_2, x_3, x_4 given as $(-1, -1, -1, -1)$ and $(1, 0, 1, 0)$, respectively, and let f_1, $f_2 \in \mathrm{Lip}_0(\mathcal{X}_0)$ be the corresponding Lipschitz functions. Then we have $\|f_1\|_L = 0.166 < 0.25 = \|f_2\|_L$ and $\|g_1\|_{V'} = 0.366 > 0.28 = \|g_2\|_{V'}$. So the norms $\|\cdot\|_{V'}$ and $\|\cdot\|_L$ do not coincide, and moreover there is no monotonic relationship between them. If the maximal margin algorithm had to choose between functions f_1 and f_2, it would come to different solutions, depending whether the underlying norm is $\|\cdot\|_{V'}$ as for the large margin classifier in V' or $\|\cdot\|_L$ as for the Lipschitz classifier in $T'V'$.

Example 7 (Normed space) *Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a normed vector space with dual $(\mathcal{X}', \|\cdot\|_{\mathcal{X}'})$. As the norm of linear functions coincides with their Lipschitz constant, $\mathcal{X}'$ is* isometrically *isomorphic to a subspace of* $\mathrm{Lip}_0(\mathcal{X}_0)$. *This means that it makes no difference whether we construct a large margin classifier on the normed space $\mathcal{X}$ directly or ignore the fact that $\mathcal{X}$ is a normed space, embed $\mathcal{X}$ into $AE(\mathcal{X}_0)$ and then construct the Lipschitz classifier on $AE(\mathcal{X}_0)$ with the subspace $T'\mathcal{X}'$. We already mentioned this fact in Section 3 when we stated that the SVM solution is the same one as the Lipschitz classifier on $\mathcal{X}'$.*

Now we want to investigate our second question: given some subspace $F \subset \mathrm{Lip}_0(\mathcal{X}_0)$, is F the dual space of some Banach space V such that $\mathcal{X}_0$ can be embedded isometrically into V and $V' \simeq F$? To answer this question we have to deal with some technical problems. First of all, F has to possess a *pre-dual*, that is a vector space V whose dual V' coincides with F. In general, not every Banach space possesses a pre-dual, and if it exists, it needs not be unique. Secondly, it turns out that the canonical injection $T' : F \to \mathrm{Lip}_0(\mathcal{X}_0)$ has to have a *pre-adjoint*, that is a mapping $T : AE(\mathcal{X}_0) \to V$ whose adjoint coincides with T'. Pre-adjoints also not always exist. In general, neither the existence of a pre-dual nor the existence of pre-adjoints are easy to prove. One situation where both can be handled is the case where F is closed under pointwise convergence:

Lemma 25 (Construction of V for given F) *Let $\mathcal{X}_0$ be a bounded metric space, and F a subspace of $(\mathrm{Lip}_0(\mathcal{X}_0), L(\cdot))$ which is closed under pointwise convergence and satisfies the condition*

$$\sup_{f \in F, L(f) \leq 1} |f(x) - f(y)| = d(x,y) \tag{15}$$

for all $x,y \in \mathcal{X}_0$. Then there exists a normed space V such that $\mathcal{X}_0$ can be isometrically embedded into V and its dual V' is isometrically isomorphic to F.

Before we can start with the proof we need two more definitions: Let M be a subspace of some Banach space V and N a subspace of the dual space V'. Then the annihilator $M^\perp$ and the pre-annihilator ${}^\perp N$ are defined as $M^\perp = \{T \in V';\ Tm = 0 \text{ for all } m \in M\}$ and ${}^\perp N = \{e \in V;\ Te = 0 \text{ for all } T \in N\}$. As the proof is a bit technical, we refer to Megginson (1998) for background reading.

Proof For a bounded metric space $\mathcal{X}_0$, the topology of pointwise convergence on $\mathrm{Lip}_0(\mathcal{X}_0)$ coincides with its weak* topology. Thus by assumption, F is weak*-closed, which implies that ${}^\perp F$ is a closed subspace of $AE(\mathcal{X}_0)$. Hence, the quotient space $V := AE(\mathcal{X}_0)/{}^\perp F$ exists, and there exists an isometric isomorphism between V' and $({}^\perp F)^\perp$. As F is weak*-closed, $({}^\perp F)^\perp = F$. So V is a pre-dual of F. Let $T' : F \to \mathrm{Lip}_0(\mathcal{X}_0)$ be the canonical inclusion. It has a pre-adjoint, namely the quotient mapping $\pi : AE(\mathcal{X}_0) \to V$. Define the mapping $\Psi : \mathcal{X}_0 \to V$, $x \mapsto \pi m_x =: v_x$. We have

$$\langle f, v_x \rangle = \langle f, \pi m_x \rangle = \langle T' f, m_x \rangle = \langle f, m_x \rangle = f(x).$$

Hence, by assumption (15), Ψ is an isometry:

$$\|\Psi(x) - \Psi(y)\|_V = \sup_{f \in F, L(f) \leq 1} \{ |\langle f, v_x - v_y \rangle| \} = \sup_{f \in F, L(f) \leq 1} \{ |f(x) - f(y)| \} = d(x,y).$$

■

Lemma 25 gives a nice interpretation of what it means geometrically to choose a subspace F of Lipschitz functions: the Lipschitz classifier with hypothesis space F corresponds to embedding $\mathcal{X}$ isometrically into the pre-dual V of F and constructing the large margin classifier on V directly. Condition (15), which F has to satisfy to allow this interpretation, intuitively means that F has to be a "reasonably large" subspace.

Example 8 (Linear combination of distance functions) *Let $\underline{F}$ be the subspace of* $\mathrm{Lip}(\mathcal{X})$ *consisting of functions of the form $f(x) = \sum_i a_i d(x_i, x) + b$, and $\overline{\underline{F}} \subset \mathrm{Lip}(\mathcal{X})$ its closure under pointwise convergence. As norm on $\overline{F}$ we take the Lipschitz constant. On $\overline{F}$, condition* (15) *is satisfied: trivially, we always have $\leq$ in* (15)*, and for given $x, y \in \mathcal{X}$, equality is reached for the function $f = d(x, \cdot)$. So we can conclude by Lemma 25 that the Lipschitz classifier on $\overline{F}$ has the geometrical interpretation explained above.*

7. Discussion

We derived a general approach to large margin classification on metric spaces which uses Lipschitz functions as decision functions. Although the Lipschitz algorithm, which implements this approach, has been derived in a rather abstract mathematical framework, it boils down to an intuitively plausible mechanism: it looks for a decision function which has a small Lipschitz constant. This agrees with the regularization principle that tries to avoid choosing functions with a high variation. The solution of the Lipschitz algorithm is well behaved as, by the representer theorems of Section 4, it can always be expressed by distance functions to training points. For some special cases, the solution corresponds to solutions of other well known algorithms, such as the support vector machine, the linear programming machine, or the 1-nearest neighbor classifier. We provide Rademacher complexity bounds for some of the involved function classes which can be used to bound the generalization error of the classifier.

In spite of all those nice properties there are several important questions which remain unanswered. To apply the Lipschitz algorithm in practice it is important to choose a suitable subspace of Lipschitz functions as hypothesis space. In Section 6 we found a geometrical explanation of what the choice of certain subspaces F means: it is equivalent to using a different isometric embedding of the metric space into some Banach space. But this explanation does not solve the question of which subspace we should choose in the end. Moreover, there exist isometric embeddings in certain Banach spaces which have no such interpretation in terms of subspaces of Lipschitz functions. For example, Hein and Bousquet (2003) studied the Kuratowski embedding of a metric space into its space of continuous functions to construct a large margin algorithm. As we explained in Example 6, the large margin classifier resulting from this embedding can be different from the Lipschitz classifier. It is an interesting question how different embeddings into different Banach spaces should be compared. One way to do this could be comparing the capacities of the induced function spaces. An interesting question in this context is to find the "smallest space" (for instance, in terms of the Rademacher complexities) in which a given data space can be embedded isometrically.

There is also a more practical problem connected to the choice of the subspace of Lipschitz functions. To implement the Lipschitz algorithm for a given subspace of Lipschitz functions, we need to know some way to efficiently compute the Lipschitz constants of the functions in the chosen subspace. For example, in case of the linear programming machine it was possible to bound the Lipschitz constants of the functions in the parameterized subspace of functions $\sum_i a_i d(x_i, \cdot) + b$ in terms of their parameters by $\sum_i |a_i|$. But in many cases, there is no obvious parametric representation of the Lipschitz constant of a class of functions. Then it is not clear how the task of minimizing the Lipschitz constant can be efficiently implemented.

An even more heretic question is whether isometric embeddings should be used at all. In our approach we adopted the point of view that a meaningful distance function between the training points is given by some external knowledge, and that we are not allowed to question it. But in practical applications it is often the case that distances are estimated by some heuristic procedure which might not give a sensible result for all the training points. In those cases the paradigm of isometric embedding might be too strong. Instead we could look for bi-Lipschitz embeddings or low distortion embeddings of the metric space into some Banach space, or even into some Hilbert space. We would then loose some (hopefully unimportant) information on the distances in the metric space, but the gain might consist in a simpler structure of the classification problem in the target space.

Finally, many people argue that for classification only “local properties” should be considered. One example is the assumption that the data lies on some low dimensional manifold in a higher dimensional space. In this case, the meaningful information consists of the intrinsic distances between points along the manifold. In small neighborhoods, those distances are close to the distances measured in the enclosing space, but for points which are far away from each other this is not true any more. In this setting it is not very useful to perform an isometric embedding of the metric space into a Banach space as the additional linear structure the Banach space imposes on the training data might be more misleading than helpful. Here a different approach has to be taken, but it is not clear how a large margin algorithm in this setting can be constructed, or even whether in this case the large margin paradigm should be applied at all.

Acknowledgments

We would like to thank Matthias Hein and Bernhard Schölkopf for helpful discussions.

Appendix A. Proof of Theorem 16

The idea of the proof of Theorem 16 is the following. Instead of bounding the Rademacher complexity on the whole set of functions $\mathcal{F}$, we first consider a maximal ε-separating subset $\mathcal{F}_\varepsilon$ of $\mathcal{F}$. This is a maximal subset such that all its points have distance at least ε to each other. To this special set we will apply the classical entropy bound of Dudley (1987):

Theorem 26 (Classical entropy bound) *For every class $\mathcal{F}$ of functions there exists a constant C such that*

$$\hat{R}_n(\mathcal{F}) \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\mathcal{F}, u, L_2(\mu_n))}\, du$$

where μ_n is the empirical distribution of the sample.

As a second step we then bound the error we make by computing the Rademacher complexity of $\mathcal{F}_\varepsilon$ instead of $\mathcal{F}$. This will lead to the additional offset of 2ε in Theorem 16. The following lemma can be found as Lemma 3.10 in Bousquet (2002) (for the definition of a separable process see also van der Vaart and Wellner 1996).

Lemma 27 (ε-separations of an empirical process) *Let $\{Z_t; t \in T\}$ be a separable stochastic process satisfying for $\lambda > 0$ the increment condition*

$$\forall s,t \in T : E\left(e^{\lambda(Z_t - Z_s)}\right) \leq e^{\lambda^2 c^2 d^2(s,t)/2}.$$

Let $\varepsilon \geq 0$ and $\delta > 0$. If $\varepsilon > 0$, let T_ε denote a maximal ε-separated subset of T and let $T_\varepsilon = T$ otherwise. Then for all t_0,

$$E\left(\sup_{t \in T_\varepsilon, d(t,t_0) \leq \delta} Z_t - Z_{t_0}\right) \leq 4\sqrt{2}c \int_{\varepsilon/4}^{\delta/2} \sqrt{\log N(T,u,d)}du.$$

To apply this lemma to the Rademacher complexity of a function class $\mathcal{F}$, we choose the index set $T = \mathcal{F}$, the fixed index $t_0 = f_0$ for some $f_0 \in \mathcal{F}$, the empirical process $Z_f = \frac{1}{n}\sum \sigma_i f(X_i)$, and $\delta \to \infty$. Note that the Rademacher complexity satisfies the increment condition of Lemma 27 with respect to the $L_2(\mu_n)$–distance with constant $c = \sqrt{n}$. Moreover, observe that $E(\sup_t Z_t - Z_{t_0}) = E(\sup_t Z_t) - E(Z_{t_0})$ and $E(Z_{t_0}) = E(\frac{1}{n}\sum \sigma_i f_0(X_i)) = 0$. Together with the symmetry of the distribution of Z_f we thus get the next lemma:

Lemma 28 (Entropy bound for ε-separations) *Let $(X_i)_{i=1,\dots,n}$ be iid training points with empirical distribution μ_n, $\mathcal{F}$ an arbitrary class of functions, and $\mathcal{F}_\varepsilon$ a maximal ε-separating subset of $\mathcal{F}$ with respect to $L_2(\mu_n)$- norm. Then*

$$E\left(\sup_{f \in \mathcal{F}_\varepsilon} \frac{1}{n} |\sum_i \sigma_i f(X_i)| \Big| X_1, \dots, X_n\right) \leq \frac{4\sqrt{2}}{\sqrt{n}} \int_{\varepsilon/4}^{\infty} \sqrt{\log N(\mathcal{F}, u, L_2(\mu_n))}\ du.$$

With this lemma we achieved that the integral over the covering numbers starts at $\varepsilon/4$ instead of 0 as it is the case in Theorem 26. The price we pay is that the supremum on the left hand side is taken over the smaller set $\mathcal{F}_\varepsilon$ instead of the whole class $\mathcal{F}$. Our next step is to bound the mistake we make by this procedure.

Lemma 29 *Let $\mathcal{F}$ be a class of functions and $\mathcal{F}_\varepsilon$ a maximal ε-separating subset of $\mathcal{F}$ with respect to $\|\cdot\|_{L_2(\mu_n)}$. Then $|R_n(\mathcal{F}) - R_n(\mathcal{F}_\varepsilon)| \leq 2\varepsilon$.*

Proof We want to bound the expression

$$|R_n(\mathcal{F}) - R_n(\mathcal{F}_\varepsilon)| = E\frac{1}{n}\left|\sup_{f \in \mathcal{F}} |\sum \sigma_i f(X_i)| - \sup_{f \in \mathcal{F}_\varepsilon} |\sum \sigma_i f(X_i)|\right|.$$

First look at the expression inside the expectation, assume that the σ_i and X_i are fixed and that $\sup_{f \in \mathcal{F}} |\sum \sigma_i f(x_i)| = |\sum \sigma_i f^*(x_i)|$ for some function f^* (if f^* does not exist we additionally have to use a limit argument). Let $f_\varepsilon \in \mathcal{F}_\varepsilon$ such that $\|f^* - f_\varepsilon\|_{L_2(\mu_n)} \leq 2\varepsilon$. Then,

$$\begin{aligned}\frac{1}{n}\left|\sup_{f \in \mathcal{F}} |\sum \sigma_i f(x_i)| - \sup_{f \in \mathcal{F}_\varepsilon} |\sum \sigma_i f(x_i)|\right| &\leq \frac{1}{n}\left||\sum \sigma_i f^*(x_i)| - |\sum \sigma_i f_\varepsilon(x_i)|\right| \\ &\leq \frac{1}{n}|\sum \sigma_i (f^*(x_i) - f_\varepsilon(x_i))| \leq \|f^* - f_\varepsilon\|_{L_1(\mu_n)} \leq \|f^* - f_\varepsilon\|_{L_2(\mu_n)} \leq 2\varepsilon.\end{aligned}$$

As this holds conditioned on all fixed values of σ_i and X_i we get the same for the expectation. This proves the lemma. ■

To prove Theorem 16 we now combine lemmas 28 and 29.

References

R. Arens and J. Eells. On embedding uniform and topological spaces. *Pacific Journal of Mathematics*, 6:397–403, 1956.

P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

K. Bennett and E. Bredensteiner. Duality and geometry in SVM classifiers. In P. Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 57–64. Morgan Kaufmann, San Francisco, 2000.

O. Bousquet. *Concentration Inequalities and Empirical Processes Theory Applied to the Analysis of Learning Algorithms*. PhD thesis, Ecole Polytechnique, 2002.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

V. Dobric and J. Yukich. Asymptotics for transportation costs in high dimensions. *Journal of Theoretical Probability*, 8(1):97–118, 1995.

R. M. Dudley. Universal Donsker classes and metric entropy. *Annals of Probability*, 15(4):1306–1326, 1987.

T. Graepel, R. Herbrich, B. Schölkopf, A. Smola, P. Bartlett, K. Müller, K. Obermayer, and R. Williamson. Classification of proximity data with LP machines. In *Proceedings of the Ninth International Conference on Artificial Neural Networks*, pages 304–309, 1999.

M. Hein and O. Bousquet. Maximal margin classification for metric spaces. In M. Warmuth B. Schölkopf, editor, *Proceedings of the 16th Annual Conference on Computational Learning Theory*, pages 72–86. Springer Verlag, Heidelberg, 2003.

A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional space. *American Mathematical Society Translations (2)*, 17:277–364, 1961.

R. Megginson. *An Introduction to Banach Space Theory*. Springer, New York, 1998.

S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152(1):37–55, 2003.

V. Pestov. Free Banach spaces and representations of topological groups. *Functional Analysis and Its Applications*, 20:70–72, 1986.

W. Rudin. *Functional Analysis*. McGraw-Hill Inc., Singapore, 2nd edition, 1991.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

J. M. Steele. *Probability theory and combinatorial optimization*, volume 69 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

M. Talagrand. The Ajtai-Komlós-Tusnády matching theorem for general measures. In *Probability in Banach spaces, 8 (Brunswick, ME, 1991)*, volume 30 of *Progress in Probability*, pages 39–54. Birkhäuser Boston, MA, 1992.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

N. Weaver. *Lipschitz algebras*. World Scientific, Singapore, 1999.

D. Zhou, B. Xiao, H. Zhou, and R. Dai. Global geometry of SVM classifiers. Technical Report 30-5-02, Institute of Automation, Chinese Academy of Sciences, 2002.