Distance-Based Classification with Lipschitz Functions

Ulrike von Luxburg and Olivier Bousquet

Max Planck Institute for Biological Cybernetics, Tübingen, Germany
{ule, olivier.bousquet}@tuebingen.mpg.de

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. Our approach leads to a general large margin algorithm for classification in metric spaces. To analyze this algorithm, we first prove a representer theorem. It states that there exists a solution which can be expressed as linear combination of distances to sets of training points. Then we analyze the Rademacher complexity of some Lipschitz function classes. The generality of the Lipschitz approach can be seen from the fact that several well-known algorithms are special cases of the Lipschitz algorithm, among them the support vector machine, the linear programming machine, and the 1-nearest neighbor classifier.

1 Introduction

Support vector machines 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.

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 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 $\|f\|^2$

corresponds to the inverse margin $\|\omega\|^2_{\mathcal{H}'}$ and the empirical error to the margin error (cf. sections 4.3 and 7 of [6]) . The benefits of these two dual viewpoints are that the regularization framework gives some intuition about the geometrical meaning of the norm $\|\cdot\|_{\mathcal{H}}$, 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 (X, d). From the regularization point of view, a convenient class of functions on a metric space is the class 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 (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 $x \in X$ will correspond to some $m_x \in \mathcal{B}$ and each Lipschitz function f on 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 $\|T_f\|$. Then we can construct a geometrical margin in $\mathcal{B}$ which allows to apply the usual large margin generalization bounds from statistical learning theory. 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

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

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 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 X is induced by a kernel), the linear programming machine (cf. [4]), 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. Furthermore, the Banach space in which we will embed X is in some sense the largest possible Banach space in which X can be embedded isometrically. This means that the Lipschitz algorithm on this space can be seen as a prototype for large margin algorithms on metric spaces. All other large margin algorithms are special cases of this general one.

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 as a vector lattice combination of the functions $d(x_i, \cdot)$ where x_i are the training points. In plain words this means that we always find solutions which are linear

combinations of distances to *sets* of training points. In section 5 we analyze the Lipschitz algorithm in terms of its Rademacher complexities. In particular, this gives valuable information about how fast the algorithm converges for different choices of subsets of Lipschitz functions.

2 Preliminaries: Lipschitz function spaces

In this section we introduce several Lipschitz function spaces and their properties. For a more detailed treatment we refer to [10].
A metric space (X, d) is a set X together with a metric d (i.e., d is non-negative, symmetric, fulfills $d(x, y) = 0 \Leftrightarrow x = y$ and the triangle inequality $d(x, y) + d(y, z) \leq d(x, z)$). A function $f : X \to \mathbb{R}$ on a metric space (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 $x, y \in X$. The smallest constant L such that this inequality holds is called the Lipschitz constant of f, denoted by $L(f)$. For Lipschitz functions f, g and scalars $a \in \mathbb{R}$ the Lipschitz constant has the properties $L(f + g) \leq L(f) + L(g)$, $L(af) \leq |a|L(f)$ 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. For a metric space (X, d) consider the set

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

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, i.e., spaces which have a finite diameter $\mathrm{diam}(X) := \sup_{x,y \in X} d(x, y)$. 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 X we choose the norm

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

as our default norm on the space $\mathrm{Lip}(X)$. It is easy to see that this indeed is a norm. One reason why it fits nicely in the learning setting is the following. Functions that are used as classifiers are supposed to take positive and negative values on the respective classes and thus satisfy $\|f\|_\infty = \sup_x |f(x)| \leq \sup_{x,y} |f(x) - f(y)| \leq \mathrm{diam}(X)L(f)$, that is $\|f\|_L = L(f)$. Hence, the norm of a classification function is determined by the quantity we use as regularizer later on. Some technical reasons for the choice of $\|\cdot\|_L$ will become clear later.

Another important space of Lipschitz functions is constructed as follows. Let (X_0, d) be a metric space with a distinguished "base point" e which is fixed in advance. Then,

$$\mathrm{Lip}_0(X_0) := \{f \in \mathrm{Lip}(X_0);\ f(e) = 0\}.$$

On this set, 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. To overcome this restriction, for a given bounded

metric space (X, d) we define a corresponding extended space $X_0 := X \cup \{e\}$ for a new base element e with the metric

$$d_{X_0}(x, y) = \begin{cases} d(x, y) & \text{for } x, y \in X \\ \operatorname{diam}(X) & \text{for } x \in X, y = e. \end{cases} \tag{1}$$

Note that $\operatorname{diam}(X_0) = \operatorname{diam}(X)$. Then we define the map

$$\psi : \operatorname{Lip}(X) \to \operatorname{Lip}_0(X_0), \quad \psi(f)(x) = \begin{cases} f(x) & \text{if } x \in X \\ 0 & \text{if } x = e \end{cases} \tag{2}$$

Obviously, ψ is bijective, and it is even an isometry: for $f_0 := \psi(f)$ we have

$$L(f_0) = \sup_{x,y \in X_0} \frac{|f_0(x) - f_0(y)|}{d_{X_0}(x, y)} = \max\{ \sup_{x,y \in X} \frac{|f(x) - f(y)|}{d(x, y)}, \sup_{x \in X} \frac{|f(x) - f(e)|}{d_{X_0}(x, e)} \} =$$
$$= \max\{L(f), \frac{\|f\|_\infty}{\operatorname{diam}(X)}\} = \|f\|_L$$

The space $(\operatorname{Lip}_0(X_0), L(\cdot))$ has some very useful duality properties. Let (X_0, d) be a metric space with distinguished base element e. A *molecule* of X_0 is a function $m : X_0 \to \mathbb{R}$ such that its support (i.e., the set where m has non-zero values) is a finite set and $\sum_{x \in X_0} m(x) = 0$. For $x, y \in X_0$ 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. We call the completion of the space of molecules with respect to $\|\cdot\|_{AE}$ the Arens-Eells space $AE(X_0)$. Denoting its dual space (i.e., the space of all continuous linear forms on $AE(X_0)$) by $AE(X_0)'$ the following theorem holds (cf. [10]).

Theorem 1. *$AE(X_0)'$ is isometrically isomorphic to* $\operatorname{Lip}_0(X_0)$.

This means that we can regard a Lipschitz function f on 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

$$\langle f, m \rangle = \sum_{x \in X_0} m(x) f(x). \tag{3}$$

It can be proved that $\|m_{xy}\|_{AE} = d(x, y)$ holds for all basic molecules m_{xy}. Hence, it is possible to embed X_0 isometrically in $AE(X_0)$ via

$$\Gamma : X_0 \to AE(X_0),\ x \mapsto m_{xe} \tag{4}$$

In this context note that the Arens-Eells space is a free Banach space over X_0. This means that we can express every map $g : X_0 \to V$ in some vector space V as a linear functional T_g on $AE(X_0)$ via $T_g m_{xe} := g(x)$. In particular, we can realize every isometric embedding g of X in some vector space V by composing Γ with the linear functional T_g. In this sense, $AE(X_0)$ is the biggest Banach space in which X can be embedded isometrically.
The norm $\|\cdot\|_{AE}$ has a nice geometrical interpretation in terms of the mass transportation problem: 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 $\sum a_{ij} d(f_i, s_j)$ where a_{ij} denotes the amount of the product transported from factory f_i to shop s_j and $d(f_i, s_j)$ the distance between them. 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 $(x_1, \ldots, x_m, y_1, \ldots, y_m)$ 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 (cf. [8]).

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 (X, d) be a metric space and $(x_i, y_i)_{i=1,\ldots,n} \subset X \times \{\pm 1\}$ some training data. In order to be able to define hyperplanes, we want to embed (X, d) into a vector space, but without loosing or changing the underlying metric structure. Our first step is to embed X by the identity mapping into the extended space X_0 as described in (1), which in turn is embedded into $AE(X_0)$ via (4). We denote the resulting composite embedding by

$$\Phi : X \to AE(X_0),\ x \mapsto m_x := m_{xe}$$

Secondly, we identify $\mathrm{Lip}(X)$ with $\mathrm{Lip}_0(X_0)$ according to (2) and then $\mathrm{Lip}_0(X_0)$ with $AE(X_0)'$ according to Theorem 1. Together this defines the map

$$\Psi : \mathrm{Lip}(X) \to AE(X_0)',\ f \mapsto T_f$$

Proposition 2. *The mappings Φ and Ψ have the following properties:*

1. *Φ is an isometric embedding of X into $AE(X_0)$: to every point $x \in X$ corresponds a molecule $m_x \in AE(X_0)$ such that $d(x,y) = \|m_x - m_y\|_{AE}$ for all $x, y \in X$.*
2. *$\mathrm{Lip}(X)$ is isometrically isomorphic to $AE(X_0)'$: to every Lipschitz function f on X corresponds an operator T_f on $AE(X_0)$ such that $\|f\|_L = \|T_f\|$ and vice versa.*
3. *It makes no difference whether we evaluate operators on the image of X in $AE(X_0)$ or apply Lipschitz functions on 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 (3). ☺

The message of this proposition is that it makes no difference whether we classify our training data on the space X with the decision function $\operatorname{sgn} f(x)$ or on $AE(X_0)$ with the hyperplane $\operatorname{sgn}(T_f m_x)$. The advantage of the latter is that there we can construct the margin of the classifier in a straightforward way: for a functional $T_f \in AE(X_0)'$ let $H_f := \{m \in AE(X_0);\ T_f m = 0\}$ be the hyperplane induced by T_f. We normalize the representation of the hyperplane such that $\min_{i=1,\dots,n} |T_f m_{x_i}| = 1$. Note that normalizing T_f is the same as normalizing f itself according to part 4 of Proposition 2. We define the margin of H_f, which we also call the margin of f, as

$$\rho := \inf_{\substack{i=1,\dots,n \\ m_h \in H_f}} \|m_{x_i} - m_h\|_{AE}.$$

Now for each training point m_{x_i} and each point m_h on the hyperplane,

$$1 \le |T_f m_{x_i}| = |T_f m_{x_i} - T_f m_h| = |T_f(m_{x_i} - m_h)| \le \|T_f\| \|m_{x_i} - m_h\|_{AE}$$

and thus $\rho \ge 1/\|T_f\| = 1/\|f\|_L$ because of part 2 of Proposition 2. If the training data are nontrivial (i.e., they contain points from both classes), then the decision function f has to take positive and negative values. Hence, $\|f\|_L = L(f)$ holds as we already explained in the last section. So we have proved the following theorem:

Theorem 3 (Margin of the Lipschitz classifier). *Let (X,d) be a metric space, $(x_i, y_i)_{i=1,\dots,n} \subset X \times \{\pm 1\}$ some training data containing points of both classes, and $f \in \mathrm{Lip}(X)$ such that $y_i f(x_i) \ge 1$ $(i = 1,\dots,n)$ and $\min_{i=1,\dots,n} |f(x_i)| = 1$. Then the margin ρ of the decision function $\operatorname{sgn} f(x)$ satisfies $\rho \ge 1/L(f)$.*

One nice aspect about the above construction is that the margin also has a geometrical meaning in the input space X itself: it is the minimal distance between the "separation surface" $S := \{s \in X;\ f(s) = 0\}$ and the training points. To see this, observe that for normalized f and $s \in S$ we have $1 \le |f(x_i) - f(s)| \le L d(x_i, s)$, and thus $d(x_i, s) \ge 1/L(f)$. Note also that

the relation between margins and Lipschitz constants in the context of normed vector spaces has already been observed in [7].
As a consequence of Theorem 3, a large margin algorithm on a metric space has to construct decision functions with small Lipschitz constant. This leads to the following optimization problem:

$$\text{minimize}_{f \in Lip(X)}\, L(f) \text{ subject to } y_i f(x_i) \geq 1,\ i = 1, \ldots, n \qquad (*)$$

We call a solution of this problem a (hard margin) Lipschitz classifier. Analogously to SVMs (e.g., [6]) we define the soft margin version of this algorithm

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

To implement this algorithm in practice we will have to choose reasonable subsets of Lipschitz functions. Consider the following special cases: if the metric on X is induced by a kernel k and we choose a classifier of the form $f(x) = \sum_i \alpha_i k(x_i, x) + b$, then the solution of the Lipschitz classifier coincides with the solution of the SVM. The reason is that the norm of a linear functional coincides with its Lipschitz constant. In the case where we choose 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 is the same as the linear programming machine (cf. [4]). 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 $\sum |a_i|$. 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 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 3, the same thus 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. [6]). 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 classifier $(*)$. To simplify the discussion, denote $\mathcal{D} := \{d(x, \cdot);\ x \in X\} \cup \{\mathbb{1}\}$ 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.

Theorem 4 (Representer theorem I). *Problem* $(*)$ *has a solution in the vector lattice spanned by* $\mathcal{D}_{\text{train}}$.

This is remarkable as the space $\mathrm{Lip}(X)$ 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 $\mathcal{D}_{\text{train}}$. Another way to state this theorem is the following:

Theorem 5 (Representer theorem II). *Problem* $(*)$ *always has a solution which is a linear combination of distances to* sets *of training points.*

To prove these theorems we first need a simple proposition. 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 $A, B \subset X$ we define $d(A,B) := \inf_{a\in A, b\in B} d(a,b)$.

Proposition 6. *The Lipschitz constant* L^* *of a solution of* $(*)$ *satisfies* $L^* \geq \frac{2}{d(X^+,X^-)}$.

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

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

Proposition 7. *Let* $L^* = \frac{2}{d(X^+,X^-)}$. *The following functions solve* $(*)$*:*

$$\begin{aligned} f_l(x) &:= \max_i (y_i - L^* d(x,x_i)) \\ f_u(x) &:= \min_i (y_i + L^* d(x,x_i)) \\ f_0(x) &:= \frac{d(x,X^-) - d(x,X^+)}{d(X^+,X^-)} \end{aligned}$$

Proof. It is easy to see that f_l,f_u, and f_0 fulfill the constraint $y_i f(x_i) \geq 1$. 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 all three functions have Lipschitz constants $\leq L^*$. Thus they are solutions of $(*)$ by Proposition 6. ☺

The functions f_l,f_u, and f_0 lie in the vector lattice spanned by $\mathcal{D}_{\text{train}}$. This proves Theorem 4. As f_0 is a linear combination of distances to sets of training points we also have proved Theorem 5.
A further remarkable fact of Proposition 7 is that the function f_0 realizes the 1-nearest neighbor classifier. This means that according to section 3 this classifier actually is a large margin classifier.
So far we have proved that $(*)$ always has a solution which can be stated 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 8. Assume four training points x_1, x_2, x_3, x_4 with (singular!) distance matrix [0 2 1 1; 2 0 1 1; 1 1 0 2; 1 1 2 0] (in matlab notation) and label vector $y = (1, 1, -1, -1)$. Then the system $f(x) = \sum_{i=1}^{4} a_i d(x_i, x) + b,\ y_i f(x_i) \geq 1$ of linear inequalities has no solution. 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 5.

This means that, in order to construct solutions for $(*)$, we are in the interesting situation that it is not enough to consider distances to single training points – we have to deal with distances to sets of training points.

5 Rademacher complexities

In this section we compute capacities of $\|\cdot\|_L$-balls of Lipschitz functions. The measures of capacity we consider are the Rademacher complexity R_n and the related maximum discrepancy $\tilde{R}_n$. Both can be used effectively to bound the generalization error of a classifier (cf. [1]). For an arbitrary class $\mathcal{F}$ of functions, they are defined as

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

where σ_i are iid Rademacher random variables (i.e., $Prob(\sigma_i = +1) = Prob(\sigma_i = -1) = 1/2$), X_i and Y_i are iid sample points according to the (unknown) sample distribution, and the expectation is taken with respect to all occurring random variables. We will describe two different ways to compute these complexities 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 (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 $\mathrm{Lip}(X)$ is the same as of the corresponding unit ball in $AE(X_0)'$. 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:

$$\sup_{\|f\|_L \leq 1} |\sum_{i=1}^{n} f(x_i) - f(y_i)| = \sup_{\|T_f\| \leq 1} |\sum_{i=1}^{n} T_f m_{x_i} - T_f m_{y_i}| =$$

$$= \sup_{\|T_f\| \leq 1} |\langle T_f, \sum_{i=1}^{n} m_{x_i} - m_{y_i} \rangle| = \|\sum_{i=1}^{n} m_{x_i y_i}\|_{AE}$$

Applying this to the definition of the maximum discrepancy immediately yields

$$\tilde{R}_n(B) = \frac{1}{n} E \| \sum_{i=1}^{n} m_{X_i Y_i} \|_{AE} \tag{5}$$

As we already explained in section 2, the norm $\| \sum_{i=1}^{n} m_{X_i Y_i} \|_{AE}$ can be interpreted as the costs of a minimal bipartite matching between $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$. To compute the right hand side of (5) 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 Y_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 $Y_1, \ldots, Y_n$ drawn iid according to some distribution P. In [11] it was has been proved that for an arbitrary distribution on the unit square of $\mathbb{R}^d$ with $d \geq 3$ we have $\lim C_n / n^{d-1/d} = c > 0$ a.s. for some constant c. The upper bound $EC_n \leq c\sqrt{n \log n}$ for arbitrary distributions on the unit square in $\mathbb{R}^2$ was presented in [9]. These results, together with equation (5), lead to the following maximum discrepancies:

Theorem 9 (Maximum discrepancy of unit ball of $\mathbf{Lip}([0,1]^d)$). *Let $X = [0,1]^d \subset \mathbb{R}^d$ with the Euclidean metric. Then the maximum discrepancy of the $\| \cdot \|_L$-unit ball B of $\mathrm{Lip}(X)$ is given by*

$$\tilde{R}_n(B) \begin{cases} = c_1 \frac{1}{\sqrt[d]{n}} & \text{if } d \geq 3 \\ \leq c_2 \frac{\sqrt{\log n}}{\sqrt{n}} & \text{if } d = 2 \end{cases}$$

for some constants c_1, c_2 which are independent of n.

Note that this gives exact results rather than upper bounds in cases where we have exact results on the bipartite matching costs. This is for example the case for cubes in $\mathbb{R}^d, d \geq 3$ as Yukich's theorem 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. The proof of this theorem can be found in the appendix.

Theorem 10 (Generalized entropy bound). *Let $\mathcal{F}$ be a class of functions and $X_1, \ldots, X_n$ iid sample points with empirical distribution μ_n. Then, for every $\varepsilon > 0$,*

$$R_n(\mathcal{F}) \leq 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\varepsilon/4}^{\infty} \sqrt{\log N(\mathcal{F}, u, \| \cdot \|_{L_2(\mu_n)})}\, du$$

To apply this theorem we need to know covering numbers of spaces of Lipschitz functions. This can be found for example in [5], pp.353–357.

Theorem 11 (Covering numbers for Lipschitz function balls). *For a totally bounded metric space (X, d) and the unit ball B of $(\mathrm{Lip}(X), \|\cdot\|_L)$,*

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

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

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

Combining Theorems 10 and 11 and using $N(\mathcal{F}, u, \|\cdot\|_{L_2(\mu_n)}) \leq N(\mathcal{F}, u, \|\cdot\|_\infty)$ now gives a bound on the Rademacher complexity of balls of $\mathrm{Lip}(X)$:

Theorem 12 (Rademacher complexity of unit ball of Lip(X)). *Let (X, d) be a totally bounded metric space with diameter $\mathrm{diam}(X)$ and B the ball of Lipschitz functions with $\|f\|_L \leq 1$. Then, for every $\varepsilon > 0$,*

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

If, in addition, X is connected and centered, we have

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

In our framework this is a nice result as the bound on the complexity of balls of $\mathrm{Lip}(X)$ only uses the metric properties of the underlying space X.

Example 13. Let $X = [0, 1]^d \subset \mathbb{R}^d, d \geq 3$, with the Euclidean metric $\|\cdot\|_2$. This is a connected and centered space. We choose $\varepsilon = 1/\sqrt[d]{n}$ and use that the covering numbers of X have the form $N(X, \varepsilon, \|\cdot\|_2) = c/\varepsilon^d$. After evaluating the second integral of Theorem 12 we find that $R_n(B)$ scales as $1/\sqrt[d]{n}$.

Example 14. Let $X = [0, 1]^2 \subset \mathbb{R}^2$ with the Euclidean metric. Applying Theorem 12 similar to Example 13 yields a bound on $R_n(B)$ that scales as $\log n/\sqrt{n}$.

In case of example 13 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 9. In case of example 14 the covering number result $\log n/\sqrt{n}$ is slightly worse than the result $\sqrt{\log(n)}/\sqrt{n}$ obtained in Theorem 9.

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

$$\mathcal{F}_{rbf} := \{f : X \to \mathbb{R} |\ f(x) = \sum_{k=1}^{l} a_k g_k(d(p_k, x)),\ g_k \in \mathcal{G}\} \tag{6}$$

where $p_k \in X$, $a_k \in \mathbb{R}$, and $\mathcal{G} \subset \mathrm{Lip}(X)$ 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

$$\begin{aligned}
\sup_{f \in B} |\sum_{i=1}^{n} \sigma_i f(x_i)| &= \sup_{\sum |a_k| L(g_k) \leq 1, p_k \in X, g_k \in \mathcal{G}} |\sum_{i=1}^{n} \sigma_i \sum_{k=1}^{l} a_k g_k(d(p_k, x_i))| \\
&= \sup_{\sum |c_k| \leq 1, p_k \in X, g_k \in \mathcal{G}} |\sum_{i=1}^{n} \sigma_i \sum_{k=1}^{l} \frac{c_k}{L(g_k)} g_k(d(p_k, x_i))| \\
&= \sup_{\sum |c_k| \leq 1, p_k \in X, g_k \in \mathcal{G}} |\sum_{k=1}^{l} c_k \sum_{i=1}^{n} \sigma_i \frac{1}{L(g_k)} g_k(d(p_k, x_i))| \\
&= \sup_{p \in X, g \in \mathcal{G}} |\sum_{i=1}^{n} \sigma_i \frac{1}{L(g)} g(d(p, x_i))| \qquad (7)
\end{aligned}$$

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 $h_{p,g}(x) := g(d(p, x_i))/L(g)$, $\mathcal{H} := \{h_{p,g};\ p \in X, g \in \mathcal{G}\}$, and $\mathcal{G}_1 := \{g/L(g);\ g \in \mathcal{G}\}$. We rewrite the right hand side of equation (7) as

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

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

Lemma 15. $N(\mathcal{H}, 2\varepsilon, \|\cdot\|_\infty) \leq N(X, \varepsilon, d) N(\mathcal{G}_1, \varepsilon, \|\cdot\|_\infty)$.

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

$$\begin{aligned}
&\|h_{p_1,g_1} - h_{p_2,g_2}\|_\infty = \sup_{x\in X} |\frac{g_1(d(p_1,x))}{L(g_1)} - \frac{g_2(d(p_2,x))}{L(g_2)}| \\
&\leq \sup_{x\in X} \left(|\frac{g_1(d(p_1,x))}{L(g_1)} - \frac{g_1(d(p_2,x))}{L(g_1)}| + |\frac{|g_1(d(p_2,x))}{L(g_1)} - \frac{g_2(d(p_2,x))}{L(g_2)}| \right) \\
&\leq \sup_{x\in X} |d(p_1,x) - d(p_2,x)| + \|\frac{g_1}{L(g_1)} - \frac{g_2}{L(g_2)}\|_\infty \\
&\leq d(p_1,p_2) + \|\frac{g_1}{L(g_1)} - \frac{g_2}{L(g_2)}\|_\infty =: d_{\mathcal{H}}(h_{p_1,g_1}, h_{p_2,g_2}) \qquad (8)
\end{aligned}$$

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

Plugging lemma 15 in Theorem 10 yields the following Rademacher complexity:

Theorem 16 (Rademacher complexity of unit ball of $\mathcal{F}_{rbf}$). *Let B the unit ball of $(\mathcal{F}_{rbf}, \|\cdot\|_{rbf})$, $\mathcal{G}_1$ the rescaled functions of $\mathcal{G}$ as defined above, and $w := \max\{\mathrm{diam}(X,d), \mathrm{diam}(\mathcal{G}_1, \|\cdot\|_\infty)\}$. Then, for every $\varepsilon > 0$,*

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

This theorem is a huge improvement compared to Theorem 12 as instead of the covering numbers we now have log-covering numbers in the integral. As an example consider the linear programming machine on $X = [0,1]^d$. Because of $\mathcal{G} = \{id\}$, the second term in the square root vanishes, and the integral over the log-covering numbers of 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 Conclusion

We derived a general approach to large margin classification on metric spaces. Our theoretical analysis led to a general algorithm that works directly on the given metric space and uses Lipschitz functions as decision functions. It specializes to well-known algorithms, as the support vector machine or the linear programming machine. Especially for the latter, our analysis gave new insights into its learning theoretic properties.

Acknowledgements

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

Appendix: Proof of Theorem 10

The idea of the proof of Theorem 10 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 [3]:

Theorem 17 (Classical entropy bound). *For every class $\mathcal{F}$ of functions there exists a constant C such that*

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

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 10. The following lemma can be found as Lemma 3.10 in [2].

Lemma 18 (ε-separations of an empirical process). *Let $\{Z_t; t \in T\}$ be a separable stochastic process satisfying for $\lambda > 0$ the increment condition*

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

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 $\delta = \infty$. Note that the Rademacher complexity satisfies the increment condition of Lemma 18 with respect to the $L_2(\mu_n)$–distance with constant $c = \sqrt{n}$. Using $E(Z_{t_0}) = E(\frac{1}{n}\sum \sigma_i f_0(X_i)) = 0$ and the symmetry of the distribution of Z_f we thus get the next lemma:

Lemma 19 (Entropy bound for ε-separations). *Let $(X_i)_{i=1,\ldots,n}$ 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*

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

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 17. 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 20. *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^* doesn't 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,

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

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

To prove Theorem 10 we now combine lemmas 19 and 20. ☺

References

1. P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.
2. O. Bousquet. Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms. PhD Thesis, 2002.
3. R. M. Dudley. Universal Donsker classes and metric entropy. *Ann. Probab.*, 15(4):1306–1326, 1987.
4. 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.
5. A. N. Kolmogorov and V. M. Tihomirov. ε-entropy and ε-capacity of sets in functional space. *Amer. Math. Soc. Transl. (2)*, 17:277–364, 1961.
6. B. Schölkopf and A. Smola. *Learning with Kernels. Support Vector Machines, Regularization, Optimization and Beyond*. MIT press, 2002.
7. B. Schölkopf, A.J. Smola, R.C. Williamson, and P.L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000.
8. J. Michael 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.
9. M. Talagrand. The Ajtai-Komlos-Tusnady matching theorem for general measures. In *Progress in Probability*, volume 30, 1991.
10. N. Weaver. *Lipschitz algebras*. World Scientific, 1999.
11. J. Yukich. Asymptotics for transportation costs in high dimensions. *J. Theor. Probab.*, 8(1):97–118, 1995.