On Kernels, Margins, and Low-dimensional Mappings

Maria-Florina Balcan[1] and Avrim Blum[1] and Santosh Vempala[2]

[1] Computer Science Department, Carnegie Mellon University
`{ninamf,avrim}@cs.cmu.edu`
[2] Department of Mathematics, MIT
`vempala@math.mit.edu`

Abstract. Kernel functions are typically viewed as providing an implicit mapping of points into a high-dimensional space, with the ability to gain much of the power of that space without incurring a high cost if data is separable in that space by a large margin γ. However, the Johnson-Lindenstrauss lemma suggests that in the presence of a large margin, a kernel function can also be viewed as a mapping to a *low*-dimensional space, one of dimension only $\tilde{O}(1/\gamma^2)$. In this paper, we explore the question of whether one can efficiently compute such implicit low-dimensional mappings, using only black-box access to a kernel function. We answer this question in the affirmative if our method is also allowed black-box access to the underlying distribution (i.e., unlabeled examples). We also give a lower bound, showing this is not possible for an arbitrary black-box kernel function, if we do not have access to the distribution. We leave open the question of whether such mappings can be found efficiently without access to the distribution for standard kernel functions such as the polynomial kernel.

Our positive result can be viewed as saying that designing a good kernel function is much like designing a good feature space. Given a kernel, by running it in a black-box manner on random unlabeled examples, we can generate an explicit set of $\tilde{O}(1/\gamma^2)$ features, such that if the data was linearly separable with margin γ under the kernel, then it is approximately separable in this new feature space.

1 Introduction

Kernels and margins have been a powerful combination in Machine Learning. A kernel function implicitly allows one to map data into a high-dimensional space and perform certain operations there without paying a high price computationally. Furthermore, if the data indeed has a large margin linear separator in that space, then one can avoid paying a high price in terms of sample size as well [6, 7, 9, 11, 13, 12, 14, 15].

The starting point for this paper is the observation that if a learning problem indeed has the large margin property under some kernel $K(x, y) = \phi(x) \cdot \phi(y)$, then by the Johnson-Lindenstrauss lemma, a *random* linear projection of the

"ϕ-space" down to a *low* dimensional space approximately preserves linear separability [1, 2, 8, 10]. Specifically, if a target function has margin γ in the ϕ-space, then a random linear projection of the ϕ-space down to a space of dimension $d = O\left(\frac{1}{\gamma^2}\log\frac{1}{\varepsilon\delta}\right)$ will, with probability at least $1-\delta$, have a linear separator of error at most ε (see, e.g., Arriaga and Vempala [2] and also Theorem 3 of this paper). This means that for any kernel K and margin γ, we can, in principle, think of K as mapping the input space X into an $\tilde{O}(1/\gamma^2)$-dimensional space, in essence serving as a representation of the data in a new (and not too large) feature space.

The question we consider in this paper is whether, given kernel K, we can in fact produce such a mapping efficiently. The problem with the above observation is that it requires explicitly computing the function $\phi(x)$. In particular, the mapping of X into R^d is a function $F(x) = A\phi(x)$, where A is a random matrix. However, for a given kernel K, the dimensionality and description of $\phi(x)$ might be large or even unknown. Instead, what we would like is an efficient procedure that given $K(.,.)$ as a black-box program, produces a mapping with the desired properties but with running time that depends (polynomially) only on $1/\gamma$ and the time to compute the kernel function K, with no dependence on the dimensionality of the ϕ-space.

Our main result is a positive answer to this question, if our procedure for computing the mapping is also given black-box access to the distribution D (i.e., unlabeled data). Specifically, given black-box access to a kernel function $K(x, y)$, a margin value γ, access to unlabeled examples from distribution D, and parameters ε and δ, we can in polynomial time construct a mapping of the feature space $F : X \to R^d$ where $d = O\left(\frac{1}{\gamma^2}\log\frac{1}{\varepsilon\delta}\right)$, such that if the target concept indeed has margin γ in the ϕ-space, then with probability $1-\delta$ (over randomization in our choice of mapping function), the induced distribution in R^d is separable with error $\leq \varepsilon$.

In particular, if we set $\varepsilon \ll \varepsilon'\gamma^2$, where ε' is our input error parameter, then the error rate of the induced target function in R^d is sufficiently small that a set S of $\tilde{O}(d/\varepsilon')$ labeled examples will, with high probability, be perfectly separable in the mapped space. This means that if the target function was truly separable with margin γ in the ϕ-space, we can apply an arbitrary zero-noise linear-separator learning algorithm in the mapped space (such as a highly-optimized linear-programming package). In fact, with high probability, not only will the data in R^d be separable, but it will be separable with margin $\gamma/2$. However, while the dimension d has a logarithmic dependence on $1/\varepsilon$, the number of (unlabeled) examples we use to produce the mapping is $\tilde{O}(1/(\gamma^2\epsilon))$.

Given the above results, a natural question is whether it might be possible to perform mappings of this type without access to the underlying distribution. In Section 5 we show that this is in general *not* possible, given only black-box access (and polynomially-many queries) to an *arbitrary* kernel K. However, it may well be possible for specific standard kernels such as the polynomial kernel or the gaussian kernel.

Our goals are to some extent related to those of Ben-David et al [4, 5]. They show negative results giving simple learning problems where one cannot construct mappings to low-dimensional spaces that preserve separability. We restrict ourselves to situations where we know that such mappings exist, but our goal is to produce them efficiently.

Outline of results: We begin in Section 3 by giving a simple mapping into a d-dimensional space for $d = O(\frac{1}{\varepsilon}[\frac{1}{\gamma^2} + \ln \frac{1}{\delta}])$ that approximately preserves both separability and margin. This mapping in fact is just the following: we draw a set S of d examples from D, run $K(x, y)$ over all pairs $x, y \in S$ to place S *exactly* into R^d, and then for general $x \in X$ define $F(x)$ to be the orthogonal projection of $\phi(x)$ down to this space (which can be computed using the kernel). That is, this mapping can be viewed as an orthogonal projection of the ϕ-space down to the space spanned by $\phi(S)$. In Section 4, we give a more sophisticated mapping to a space of dimension only $O(\frac{1}{\gamma^2} \log \frac{1}{\varepsilon\delta})$. This logarithmic dependence then means we can set ε small enough as a function of the dimension and our input error parameter that we can then plug in a generic zero-noise linear separator algorithm in the mapped space (assuming the target function was perfectly separable with margin γ in the ϕ-space). In Section 5 we argue that for a black-box kernel, one must have access to the underlying distribution D if one wishes to produce a good mapping into a low-dimensional space. Finally, we give a short discussion in Section 6.

An especially simple mapping: We also note that a corollary to one of our results (Lemma 1) is that if we are willing to use dimension $d = O(\frac{1}{\varepsilon}[\frac{1}{\gamma^2} + \ln \frac{1}{\delta}])$ and we are not concerned with preserving the margin and only want approximate separability, then the following especially simple procedure suffices. Just draw a random sample of d unlabeled points $x_1, \ldots, x_d$ and define $F(x) = (K(x, x_1), \ldots, K(x, x_d))$. That is, we define the ith "feature" of x to be $K(x, x_i)$. Then, with high probability, the data will be approximately separable in this d-dimensional space if the target function had margin γ in the ϕ space. Thus, this gives a particularly simple way of using the kernel and distribution for feature generation.

2 Notation and Definitions

We assume that data is drawn from some distribution D over an instance space X and labeled by some unknown target function $c : X \to \{-1, +1\}$. We use P to denote the combined distribution over labeled examples.

A *kernel* K is a pairwise function $K(x, y)$ that can be viewed as a "legal" definition of inner product. Specifically, there must exist a function ϕ mapping X into a possibly high-dimensional Euclidean space such that $K(x, y) = \phi(x) \cdot \phi(y)$. We call the range of ϕ the "ϕ-space", and use $\phi(D)$ to denote the induced distribution in the ϕ-space produced by choosing random x from D and then applying $\phi(x)$.

We say that for a set S of labeled examples, a vector w in the ϕ-space has margin γ if:

$$\min_{(x,\ell)\in S}\left[\ell\frac{w\cdot\phi(x)}{|w||\phi(x)|}\right]\geq\gamma.$$

That is, w has margin γ if any labeled example in S is correctly classified by the linear separator $w\cdot\phi(x)\geq 0$, and furthermore the cosine of the angle between w and $\phi(x)$ has magnitude at least γ. If such a vector w exists, then we say that S is linearly separable with margin γ under the kernel K. For simplicity, we are only considering separators that pass through the origin, though our results can be adapted to the general case as well.

We can similarly talk in terms of the distribution P rather than a sample S. We say that a vector w in the ϕ-space has margin γ with respect to P if:

$$\Pr_{(x,\ell)\in P}\left[\ell\frac{w\cdot\phi(x)}{|w||\phi(x)|}<\gamma\right]=0.$$

If such a vector w exists, then we say that P is (perfectly) linearly separable with margin γ under K. One can also weaken the notion of perfect separability. We say that a vector w in the ϕ-space has error α at margin γ if:

$$\Pr_{(x,\ell)\in P}\left[\ell\frac{w\cdot\phi(x)}{|w||\phi(x)|}<\gamma\right]\leq\alpha.$$

Our starting assumption in this paper will be that P is perfectly separable with margin γ under K, but we can also weaken the assumption to the existence of a vector w with error α at margin γ, with a corresponding weakening of the implications. Our goal is a mapping $F: X\to R^d$ where d is not too large that approximately preserves separability. We use $F(D)$ to denote the induced distribution in R^d produced by selecting points in X from D and then applying F, and use $F(P)=F(D,c)$ to denote the induced distribution on labeled examples.

For a set of vectors $v_1,v_2,\ldots,v_k$ in Euclidean space, let $\mathrm{span}(v_1,\ldots,v_k)$ denote the span of these vectors: that is, the set of vectors v that can be written as a linear combination $a_1v_1+\ldots+a_kv_k$. Also, for a vector v and a subspace Y, let $\mathrm{proj}(v,Y)$ be the orthogonal projection of v down to Y. So, for instance, $\mathrm{proj}(v,\mathrm{span}(v_1,\ldots,v_k))$ is the orthogonal projection of v down to the space spanned by $v_1,\ldots,v_k$. We note that given a set of vectors $v_1,\ldots,v_k$ and the ability to compute dot-products, this projection can be computed efficiently by a solving a set of linear equalities.

3 A simpler mapping

Our goal is a procedure that given black-box access to a kernel function $K(.,.)$, unlabeled examples from distribution D, and a margin value γ, produces a (probability distribution over) mappings $F: X\to R^d$ such that if the target function indeed has margin γ in the ϕ-space, then with high probability our mapping will preserve approximate linear separability. In this section, we analyze a method that produces a space of dimension $O\left(\frac{1}{\varepsilon}\left[\frac{1}{\gamma^2}+\ln\frac{1}{\delta}\right]\right)$, where ε is our bound on the error rate of the best separator in the mapped space. We will, in fact, strengthen our goal somewhat to require that $F(P)$ be approximately separable at margin $\gamma/2$ (rather than just approximately separable) so that we can use this mapping as a first step in a better mapping in Section 4.

Informally, the method is just to draw a set S of d examples from D, and then (using the kernel K) to define $F(x)$ so that it is equivalent to an orthogonal projection of $\phi(x)$ down to the space spanned by $\phi(S)$.

The following lemma is key to our analysis.

Lemma 1. *Consider any distribution over labeled examples in Euclidean space such that there exists a vector w with margin γ. Then if we draw*

$$n \geq \frac{8}{\varepsilon}\left[\frac{1}{\gamma^2}+\ln\frac{1}{\delta}\right]$$

examples $z_1,\ldots,z_n$ iid from this distribution, with probability $\geq 1-\delta$, there exists a vector w' in $\mathrm{span}(z_1,\ldots,z_n)$ that has error at most ε at margin $\gamma/2$.

Proof. We give here two proofs of this lemma. The first (which produces a somewhat worse bound on n) uses the machinery of margin bounds. Margin bounds [12, 3] tell us that using $n = O(\frac{1}{\varepsilon}[\frac{1}{\gamma^2}\log^2(1/\gamma\varepsilon)+\log\frac{1}{\delta}])$ points, with high probability, *any* separator with margin $\geq \gamma$ over the observed data has a low true error rate. Thus, the projection of the target function w into this space will have a low error rate as well. (Projecting w into this space maintains the value of $w \cdot z_i$, while possibly shrinking the vector w, which can only increase the margin over the observed data.) The only technical issue is that we want as a conclusion for the separator not only to have a low error rate over the distribution, but also to have a large margin. However, we can easily get this from the standard double-sample argument. Specifically, rather than use a $\gamma/2$-cover as in the standard margin bound, one can use a $\gamma/4$-cover. When the double sample is randomly partitioned into (S_1, S_2), it is unlikely that any member of this cover will have zero error on S_1 at margin $3\gamma/4$, and yet substantial error on S_2 at the same margin, which then implies that (since this is a $\gamma/4$-cover) no separator has zero error on S_1 at margin γ and yet substantial error on S_2 at margin $\gamma/2$.

However, we also note that since we are only asking for an existential statement (the *existence* of w'), we do not need the full machinery of margin bounds, and give a second more direct proof (with better bounds on n) from first principles. For any set of points S, let $w_{in}(S)$ be the projection of w to $\mathrm{span}(S)$, and let $w_{out}(S)$ be the orthogonal portion of w, so that $w = w_{in}(S) + w_{out}(S)$ and $w_{in}(S) \perp w_{out}(S)$. Also, for convenience, assume w and all examples z are unit-length vectors (since we have defined margins in terms of angles, we can do this without loss of generality). Now, let us make the following definitions. Say that $w_{out}(S)$ is *large* if $\Pr_z(|w_{out}(S)\cdot z| > \gamma/2) \geq \varepsilon$, and otherwise say that $w_{out}(S)$ is *small*. Notice that if $w_{out}(S)$ is small, we are done, because $w \cdot z = (w_{in}(S)\cdot z) + (w_{out}(S)\cdot z)$, which means that $w_{in}(S)$ has the properties we want. On the other hand, if $w_{out}(S)$ is large, this means that a new random point z has at least an ε chance of improving the set S. Specifically, consider z such that $|w_{out}(S) \cdot z| > \gamma/2$. For $S' = S \bigcup \{z\}$, we have $w_{out}(S') = w_{out}(S) - \text{proj}(w_{out}(S), \text{span}(S')) = w_{out}(S) - (w_{out}(S) \cdot z')z'$, where $z' = (z - \text{proj}(z, \text{span}(S)))/|z - \text{proj}(z, \text{span}(S))|$ is the portion of z orthogonal to $\text{span}(S)$, stretched to be a unit vector. But since $|w_{out}(S) \cdot z'| \geq |w_{out}(S) \cdot z|$, this implies that $|w_{out}(S')|^2 < |w_{out}(S)|^2 - (\gamma/2)^2$. Now, since $|w|^2 = |w_{out}(\emptyset)|^2 = 1$ and $|w_{out}(S)|$ can never become negative, this can happen at most $4/\gamma^2$ times. So, we have a situation where so long as w_{out} is large, each example has at least an ε chance of reducing $|w_{out}|^2$ by at least $\gamma^2/4$. This can happen at most $4/\gamma^2$ times, so Chernoff bounds imply that with probability at least $1 - \delta$, $w_{out}(S)$ will be small for S a sample of size $\geq \frac{8}{\varepsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]$. □

Lemma 1 implies that if P is linearly separable with margin γ under K, and we draw $n = \frac{8}{\varepsilon}[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}]$ random unlabeled examples $x_1, \ldots, x_n$ from D, with probability at least $1 - \delta$ there is a separator w' in the ϕ-space with error rate at most ϵ that can be written as

$$w' = \alpha_1\phi(x_1) + \ldots + \alpha_n\phi(x_n).$$

Notice that since $w' \cdot \phi(x) = \alpha_1 K(x, x_1) + \ldots + \alpha_n K(x, x_n)$, an immediate implication is that if we simply think of $K(x, x_i)$ as the ith "feature" of x — that is, if we define $\hat{F}(x) = (K(x, x_1), \ldots, K(x, x_n))$ — then with high probability $\hat{F}(P)$ will be approximately linearly separable as well. So, the kernel and distribution together give us a particularly simple way of performing feature generation that preserves (approximate) separability.

Unfortunately, the above mapping $\hat{F}$ may not preserve margins because we do not have a good bound on the length of the vector $(\alpha_1, \ldots, \alpha_n)$ defining the separator in the new space. Instead, to preserve margin we want to perform an orthogonal projection. Specifically, we draw a set $S = \{x_1, ..., x_n\}$ of $\frac{8}{\varepsilon}[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}])$ unlabeled examples from D and run $K(x, y)$ for all pairs $x, y \in S$. Let $M(S) = (K(x_i, x_j))_{x_i, x_j \in S}$ be the resulting kernel matrix. We use $M(S)$ to define an embedding of S into R^n by Cholesky Factorization. More specifically, we decompose $M(S)$ into $M(S) = U'U$, where U is an upper triangular matrix, and we define our mapping $F(x_j)$ to be the j'th column of U.

We next extend the embedding to all of X by considering $F : X \to R^n$ to be a mapping defined as follows: for $x \in X$, let $F(x) \in R^n$ be the point such that $F(x) \cdot F(x_i) = K(x, x_i)$, for all $i \in \{1, ..., n\}$. In other words, this mapping is equivalent to orthogonally projecting $\phi(x)$ down to $\text{span}(\phi(x_1), \ldots, \phi(x_n))$. We can compute $F(x)$ by solving the system of linear equations $[F(x)]'U = (K(x, x_1), ..., K(x, x_n))$.

We now claim that by Lemma 1, this mapping F maintains approximate separability at margin $\gamma/2$.

Theorem 1. *Given $\varepsilon, \delta, \gamma < 1$, if P has margin γ in the ϕ-space, then with probability $\geq 1 - \delta$ our mapping F (into the space of dimension n) has the* *property that $F(P)$ is linearly separable with error at most ε at margin $\gamma/2$, given that we use $n \geq \frac{8}{\varepsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]$ unlabeled examples.*

Proof. Since $\phi(D)$ is separable at margin γ, it follows from Lemma 1 that, for $n \geq \frac{8}{\varepsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]$, with probability at least $1-\delta$, there exists a vector w that can be written as $w = \alpha_1\phi(x_1) + ... + \alpha_n\phi(x_n)$, that has error at most ε at margin $\gamma/2$ (with respect to $\phi(P)$), i.e.,

$$\Pr_{(x,\ell)\in P}\left[\frac{\ell(w\cdot\phi(x))}{|w||\phi(x)|} < \frac{\gamma}{2}\right] \leq \varepsilon.$$

Consider $\overline{w} \in R^n$, $\overline{w} = \alpha_1 F(x_1) + ... + \alpha_n F(x_n)$. Since $|\overline{w}| = |w|$ and since $w\cdot\phi(x) = \overline{w}\cdot F(x)$ and $|F(x)| \leq |\phi(x)|$ for every $x \in X$, we get that $\overline{w}$ has error at most ε at margin $\gamma/2$ (with respect to $F(P)$), i.e.,

$$\Pr_{(x,\ell)\in P}\left[\frac{\ell(\overline{w}\cdot F(x))}{|\overline{w}||F(x)|} < \frac{\gamma}{2}\right] \leq \varepsilon.$$

Therefore, for our choice of n, with probability at least $1-\delta$ (over randomization in our choice of F), there exists a vector $\overline{w} \in R^n$ that has error at most ε at margin $\gamma/2$ with respect to $F(P)$. □

Notice that the running time to compute $F(x)$ is polynomial in $1/\gamma, 1/\epsilon, 1/\delta$ and the time to compute the kernel function K.

4 An improved mapping

We now describe an improved mapping, in which the dimension d has only a logarithmic, rather than linear, dependence on $1/\varepsilon$. The idea is to perform a two-stage process, composing the mapping from the previous section with a random linear projection from the range of that mapping down to the desired space. Thus, this mapping can be thought of as combining two types of random projection: a projection based on points chosen at random from D, and a projection based on choosing points uniformly at random in the intermediate space.

We begin by stating a result from [2] that we will use. Here $N(0,1)$ is the standard Normal distribution with mean 0 and variance 1 and $U(-1,1)$ is the distribution that has probability 1/2 on -1 and probability 1/2 on 1.

Theorem 2 (Neuronal RP [2]). *Let $u, v \in R^n$. Let $u' = \frac{1}{\sqrt{k}}Au$ and $v' = \frac{1}{\sqrt{k}}Av$ where A is a random matrix whose entries are chosen independently from either $N(0,1)$ or $U(-1,1)$. Then,*

$$\Pr_A\left[(1-\varepsilon)|u-v|^2 \leq |u'-v'|^2 \leq (1+\varepsilon)|u-v|^2\right] \geq 1 - 2e^{-(\varepsilon^2-\varepsilon^3)\frac{k}{4}}.$$

Let $F_1 : X \to R^n$ be the mapping from Section 3, with $\varepsilon/2$ and $\delta/2$ as its error and confidence parameters respectively. Let $F_2 : R^n \to R^d$ be a random projection as in Theorem 2. Specifically, we pick A to be a random $d \times n$ matrix whose entries are chosen i.i.d. $N(0,1)$ or $U(-1,1)$ (i.e., uniformly from $\{-1,1\}$). We then set $F_2(x) = \frac{1}{\sqrt{d}}Ax$. We finally consider our overall mapping $F : X \to R^d$ to be $F(x) = F_2(F_1(x))$.

We now claim that for $n = O\left(\frac{1}{\varepsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]\right)$ and $d = O\left(\frac{1}{\gamma^2}\log(\frac{1}{\varepsilon\delta})\right)$, with high probability, this mapping has the desired properties. The basic argument is that the initial mapping F_1 maintains approximate separability at margin $\gamma/2$ by Lemma 1, and then the second mapping approximately preserves this property by Theorem 2.

Theorem 3. *Given $\varepsilon, \delta, \gamma < 1$, if P has margin γ in the ϕ-space, then with probability at least $1-\delta$, our mapping into the space of dimension $d = O\left(\frac{1}{\gamma^2}\log(\frac{1}{\varepsilon\delta})\right)$ has the property that $F(P)$ is linearly separable with error at most ε at margin at most $\gamma/4$, given that we use $n = O\left(\frac{1}{\varepsilon}\left[\frac{1}{\gamma^2} + \ln\frac{1}{\delta}\right]\right)$ unlabeled examples.*

Proof. By Lemma 1, with probability at least $1-\delta/2$ there exists a separator w in the intermediate space R^n with error at most $\varepsilon/2$ at margin $\gamma/2$. Let us assume this in fact occurs. Now, consider some point $x \in R^n$. Theorem 2 implies that under the random projection F_2, with high probability the lengths of w, x, and $w-x$ are all approximately preserved, which implies that the cosine of the angle between w and x (i.e., the margin of x with respect to w) is also approximately preserved. Specifically, for $d = O\left(\frac{1}{\gamma^2}\log(\frac{1}{\varepsilon\delta})\right)$, we have:

$$\text{For all } x, \quad \Pr_A\left[\left|\frac{w \cdot x}{|w||x|} - \frac{F_2(w)\cdot F_2(x)}{|F_2(w)||F_2(x)|}\right| \geq \gamma/4\right] \leq \epsilon\delta/4.$$

This implies

$$\Pr_{x \in F_1(D), A}\left[\left|\frac{w \cdot x}{|w||x|} - \frac{F_2(w)\cdot F_2(x)}{|F_2(w)||F_2(x)|}\right| \geq \gamma/4\right] \leq \epsilon\delta/4,$$

which implies that

$$\Pr_A\left[\Pr_{x \in F_1(D)}\left(\left|\frac{w \cdot x}{|w||x|} - \frac{F_2(w)\cdot F_2(x)}{|F_2(w)||F_2(x)|}\right| \geq \gamma/4\right) \geq \epsilon/2\right] \leq \delta/2.$$

Since w has error $\leq \varepsilon/2$ at margin $\gamma/2$, this then implies that the probability that $F_2(w)$ has error more than ε over $F_2(F_1(D))$ at margin $\gamma/4$ is at most $\delta/2$. Combining this with the $\delta/2$ failure probability of F_1 completes the proof. □

As before, the running time to compute our mappings is polynomial in $1/\gamma, 1/\epsilon, 1/\delta$ and the time to compute the kernel function K.

Corollary 1. *Given $\varepsilon', \delta, \gamma < 1$, if P has margin γ in the ϕ-space then we can use $n = \tilde{O}(1/(\epsilon'\gamma^4))$ unlabeled examples to produce a mapping into R^d for $d = O(\frac{1}{\gamma^2}\log\frac{1}{\epsilon'\gamma\delta})$, that with probability $1-\delta$ has the property that $F(P)$ is linearly separable with error $\ll \epsilon'/d$.*

Proof. Just plug in the desired error rate into the bounds of Theorem 3. □

Note that we can set the error rate in Corollary 1 so that with high probability a random labeled set of size $\tilde{O}(d/\epsilon')$ will be linearly separable, and therefore any linear separator will have low error by standard VC-dimension arguments. Thus, we can apply an arbitrary linear-separator learning algorithm in R^d to learn the target concept.

5 On the necessity of access to D

Our main algorithm constructs a mapping $F : X \to R^d$ using black-box access to the kernel function $K(x, y)$ together with unlabeled examples from the input distribution D. It is natural to ask whether it might be possible to remove the need for access to D. In particular, notice that the mapping resulting from the Johnson-Lindenstrauss lemma has nothing to do with the input distribution: if we have access to the ϕ-space, then no matter what the distribution is, a random projection down to R^d will approximately preserve the existence of a large-margin separator with high probability. So perhaps such a mapping F can be produced by just computing K on some polynomial number of cleverly-chosen (or uniform random) points in X.[3] In this section, we give an argument showing why this may not be possible for an arbitrary kernel. This leaves open, however, the case of specific natural kernels.

In particular, consider $X = \{0, 1\}^n$, let X' be a random subset of $2^{n/2}$ elements of X, and let D be the uniform distribution on X'. For a given target function c, we will define a special ϕ-function ϕ_c such that c is a large margin separator in the ϕ-space under distribution D, but that only the points in X' behave nicely, and points not in X' provide no useful information. Specifically, consider $\phi_c : X \to R^2$ defined as:

$$\phi_c(x) = \begin{cases} (1, 0) & \text{if } x \notin X' \\ (-1/2, \sqrt{3}/2) & \text{if } x \in X' \text{ and } c(x) = 1 \\ (-1/2, -\sqrt{3}/2) & \text{if } x \in X' \text{ and } c(x) = -1 \end{cases}$$

See figure 1. This then induces the kernel:

$$K_c(x, y) = \begin{cases} 1 & \text{if } x, y \notin X' \text{ or } [x, y \in X' \text{ and } c(x) = c(y)] \\ -1/2 & \text{otherwise} \end{cases}$$

Notice that the distribution $P = (D, c)$ over labeled examples has margin $\gamma = 1/2$ in the ϕ-space.

Now, consider any algorithm with black-box access to K attempting to create a mapping $F : X \to R^d$. Since X' is a random exponentially-small fraction of X, with high probability all calls made to K return the value 1. Furthermore, even though at "runtime" when x is chosen from D, the function $F(x)$ may itself call

[3] Let's assume X is a "nice" space such as the unit ball or $\{0, 1\}^n$.

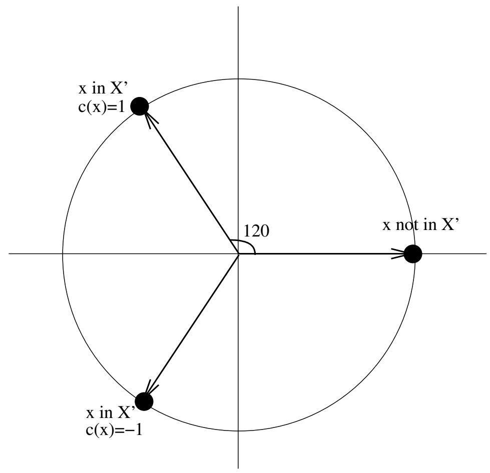

Fig. 1. Function ϕ_c used in lower bound.

$K(x, y)$ for different previously-computed points y, with high probability these will all give $K(x, y) = -1/2$. In particular, this means that the mapping F is with high probability independent of the target function c. Now, since X' has size $2^{n/2}$, there are exponentially many orthogonal functions c over D, which means that with high probability $F(D, c)$ will not even be weakly separable for a random function c over X' unless d is exponentially large in n.

Notice that the kernel in the above argument is positive semidefinite. If we wish to have a positive definite kernel, we can simply change "1" to "$1-\alpha$" and "$-1/2$" to "$-\frac{1}{2}(1-\alpha)$" in the definition of $K(x, y)$, except for $y = x$ in which case we keep $K(x, y) = 1$. This corresponds to a function ϕ in which rather that mapping points exactly into R^2, we map into R^{2+2^n} giving each example a $\sqrt{\alpha}$-component in its own dimension, and we scale the first two components by $\sqrt{1-\alpha}$ to keep $\phi_c(x)$ a unit vector. The margin now becomes $\frac{1}{2}(1-\alpha)$. Since the modifications provide no real change (an algorithm with access to the original kernel can simulate this one), the above arguments apply to this kernel as well.

Of course, these kernels are extremely unnatural, each with its own hidden target function built in. It seems quite conceivable that positive results independent of the distribution D can be achieved for standard, natural kernels.

6 Discussion and Open Problems

Our results show that given black-box access to a kernel function K and a distribution D (i.e., unlabeled examples) we can use K and D together to construct a new low-dimensional feature space in which to place our data that approximately preserves the desired properties of the kernel.

We note that if one has an unkernelized algorithm for learning linear separators with good margin-based sample-complexity bounds, then one does not necessarily need to perform a mapping first and instead can apply a more direct method. Specifically, draw a sufficiently large *labeled* set S as required by the algorithm's sample-complexity requirements, compute the kernel matrix $K(x, y)$ to place S into $R^{|S|}$, and use the learning algorithm to find a separator h in that space. New examples can be projected into that space using the kernel function (as in Section 3) and classified by h. Thus, our result is perhaps something more of interest from a conceptual point of view, or something we could apply if one had a generic (e.g., non-margin-dependent) linear separator algorithm.

One aspect that we find conceptually interesting is the relation of the two types of "random" mappings used in our approach. On the one hand, we have mappings based on random examples drawn from D, and on the other hand we have mappings based on uniform (or Gaussian) random vectors in the ϕ-space as in the Johnson-Lindenstrauss lemma.

Our main open question is whether, for natural standard kernel functions, one can produce mappings $F : X \to R^d$ in an oblivious manner, without using examples from the data distribution. The Johnson-Lindenstrauss lemma tells us that such mappings exist, but the goal is to produce them without explicitly computing the ϕ-function.

Acknowledgements

We would like to thank Adam Kalai and John Langford for helpful discussions. This work was supported in part by NSF grants CCR-0105488, NSF-ITR CCR-0122581, and NSF-ITR IIS-0312814.

References

1. D. Achlioptas, "Database-friendly Random Projections", Symposium on Principles of Database Systems, 2001.
2. R. I. Arriaga, S. Vempala, "An algorithmic theory of learning, Robust concepts and random projection", Proc. of the 40th Foundations of Computer Science, 616 - 623, 1999.
3. P. Bartlett, J. Shawe-Taylor, "Generalization Performance of Support Vector Machines and Other Pattern Classifiers", Advances in Kernel Methods: Support Vector Learning, MIT Press, 1999.
4. S. Ben-David, N. Eiron, H.U. Simon, "Limitations of Learning Via Embeddings in Euclidean Half-Spaces", Journal of Machine Learning Research 3: 441-461, 2002.
5. S. Ben-David, "A Priori Generalization Bounds for Kernel Based Learning", NIPS 2001 Workshop on Kernel Based Learning.
6. B. E. Boser, I. M. Guyon, V. N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers", Proceedings of the Fifth Annual Workshop on Computational Learning Theory, 1992.
7. C. Cortes, V. Vapnik, "Support-Vector Networks", Machine Learning, Volume 20(3): 273 - 297, 1995.
8. S. Dasgupta, A. Gupta, "An elementary proof of the Johnson-Lindenstrauss Lemma", Tech Report, UC Berkeley, 1999.

9. Y. Freund, R. E. Schapire, "Large Margin Classification Using the Perceptron Algorithm", Machine Learning, Volume 37, No. 3, 277-296, 1999.
10. W. B. Johnson, J. Lindenstrauss, " Extensions of Lipschitz mappings into a Hilbert space", Conference in modern analysis and probability, 189–206, 1984.
11. K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, "An Introduction to kernel-based learning algorithms", IEEE Transactions on Neural Networks, Vol. 12, pp. 181-201, 2001.
12. J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, M. Anthony, "Structural Risk Minimization over Data-Dependent Hierarchies", IEEE Trans. on Information Theory, 44(5):1926-1940, 1998.
13. A. J. Smola, P. Bartlett, B. Scholkopf, D. Schuurmans (Eds.), "Advances in Large Margin Classifiers", MIT Press, 2000.
14. B. Scholkopf, A. J. Smola , "Learning with kernels. Support Vector Machines, Regularization, Optimization, and Beyond", MIT University Press, Cambridge, 2002.
15. V. N. Vapnik, "Statistical Learning Theory", John Wiley and Sons Inc., New York, 1998.