Improved Guarantees for Learning via Similarity Functions

Maria-Florina Balcan **Avrim Blum**
Computer Science Department, Carnegie Mellon University
{ninamf,avrim}@cs.cmu.edu

Nathan Srebro
Toyota Technological Institute at Chicago
nati@uchicago.edu

Abstract

We continue the investigation of natural conditions for a similarity function to allow learning, without requiring the similarity function to be a valid kernel, or referring to an implicit high-dimensional space. We provide a new notion of a "good similarity function" that builds upon the previous definition of Balcan and Blum (2006) but improves on it in two important ways. First, as with the previous definition, any large-margin kernel is also a good similarity function in our sense, but the translation now results in a much milder increase in the labeled sample complexity. Second, we prove that for distribution-specific PAC learning, our new notion is strictly more powerful than the traditional notion of a large-margin kernel. In particular, we show that for any hypothesis class C there exists a similarity function under our definition allowing learning with $O(\log |C|)$ labeled examples. However, in a lower bound which may be of independent interest, we show that for any class C of pairwise uncorrelated functions, there is *no* kernel with margin $\gamma \geq 8/\sqrt{|C|}$ for all $f \in C$, even if one allows average hinge-loss as large as 0.5. Thus, the sample complexity for learning such classes with SVMs is $\Omega(|C|)$. This extends work of Ben-David et al. (2003) and Forster and Simon (2006) who give hardness results with comparable margin bounds, but at much lower error rates.

Our new notion of similarity relies upon L_1 regularized learning, and our separation result is related to a separation result between what is learnable with L_1 vs. L_2 regularization.

1 Introduction

Kernel functions have become an extremely popular tool in machine learning, with an attractive theory as well (Scholkopf & Smola, 2002; Herbrich, 2002; Shawe-Taylor & Cristianini, 2004; Scholkopf et al., 2004). This theory views a kernel as implicitly mapping data points into a possibly very high dimensional space, and describes a kernel function as being good for a given learning problem if data is separable by a large margin in that implicit space. However, while quite elegant, this theory does not necessarily correspond to the intuition of a good kernel as a good measure of similarity, and the underlying margin in the implicit space usually is not apparent in "natural" representations of the data. Therefore, it may be difficult for a domain expert to use the theory to help design an appropriate kernel for the learning task at hand. Moreover, the requirement of positive semi-definiteness may rule out the most natural pairwise similarity functions for the given problem domain.

In recent work, Balcan and Blum (2006) developed an alternative, more general theory of learning with pairwise similarity functions that may not necessarily be valid positive semi-definite kernels. Specifically, this work developed sufficient conditions for a similarity function to allow one to learn well) that does not require reference to implicit spaces, and does not require the function to be positive semi-definite (or even symmetric). While this theory provably generalizes the standard theory in that any good kernel function in the usual sense can be shown to also be a good similarity function under this definition, the translation does incur a penalty. Subsequently, Srebro (2007) tightly quantified the gap between the learning guarantees based on kernel-based learning, and those that can be obtained by using the kernel as a similarity function in this way. In particular, Srebro (2007) shows that a kernel of margin γ is guaranteed to be a similarity function of margin $\Omega(\epsilon\gamma^2)$ at hinge-loss ϵ, and furthermore there exist examples for which this is tight. To sum up, while the theory of Balcan and Blum (2006) applies to a wider class of pairwise functions than the standard notion of kernel learning, it might be quantitatively inferior in those cases that both notions apply.

In this work we develop a new notion of a good similarity function that broadens the definition of Balcan and Blum (2006) while still guaranteeing learnability. As with the previous definition, our notion talks in terms of natural similarity-based properties and does not require positive semi-definiteness or reference to implicit spaces. However, our new notion improves on the previous definition in two important respects:

First, our new notion provides a better kernel-to-similarity translation. Any large-margin kernel function is a good similarity function under our definition, and while we still incur some loss in the parameters, this loss is much smaller than under the prior definition, especially in terms of the final labeled sample-complexity bounds. In particular, when using

a valid kernel function as a similarity function, a substantial portion of the previous sample-complexity bound can be transferred over to merely a need for *un*labeled examples.

Second, we show that our new definition allows for good similarity functions to exist for concept classes for which there is *no* good kernel. In particular, for any concept class C and sufficiently unconcentrated distribution D, we show there exists a similarity function under our definition with parameters yielding a labeled sample complexity bound of $O(\frac{1}{\epsilon}\log|C|)$ to achieve error ϵ, matching the ideal sample complexity for a generic hypothesis class. In fact, we also extend this result to classes of finite VC-dimension rather than finite cardinality. In contrast, we show there exist classes C such that under the uniform distribution over the instance space, there is no kernel with margin $8/\sqrt{|C|}$ for all $f \in C$ even if one allows 0.5 average hinge-loss. Thus, the margin-based guarantee on sample complexity for learning such classes with kernels is $\Omega(|C|)$. This extends work of Ben-David et al. (2003) and Forster and Simon (2006) who give hardness results with comparable margin bounds, but at much lower error rates. Warmuth and Vishwanathan (2005) provide lower bounds for kernels with similar error rates, but their results hold only for regression (not hinge loss). Note that given access to unlabeled data, any similarity function under the definition of Balcan and Blum (2006) can be converted to a kernel function with approximately the same parameters. Thus, our lower bound for kernel functions applies to that definition as well. These results establish a gap in the representational power of similarity functions under our new definition relative to the representational power of either kernels or similarity functions under the old definition.

Both our new definition and that of Balcan and Blum (2006) are based on the idea of a similarity function being good for a learning problem if there exists a non-negligible subset R of "reasonable points" such that most examples x are on average more similar to the reasonable points of their own label than to the reasonable points of the other label. (Formally, the "reasonableness" of an example may be given by a weight between 0 and 1 and viewed as probabilistic or fractional.) However, the previous definition combined the two quantities of interest—the probability mass of reasonable points and the gap in average similarity to reasonable points of each label—into a single margin parameter. The new notion keeps these quantities distinct, which turns out to make a substantial difference both in terms of broadness of applicability and in terms of the labeled sample complexity bounds that result.

Note that we distinguish between labeled and unlabeled sample complexities: while the total number of examples needed depends polynomially on the two quantities of interest, the number of labeled examples depends only logarithmically on the probability mass of the reasonable set and therefore may be much smaller under the new definition. This is especially beneficial in situations in which unlabeled data is plentiful (or the distribution is known and so unlabeled data is free), but labeled data is scarce.

Another way to view the distinction between the two notions of similarity is that we now require good predictions using a weight function with bounded expectation, rather than bounded supremum: compare the old Definition 4 and the variant of the new definition given as Definition 17. (We do in fact still have a bound on the supremum, but this bound only affects the labeled sampled complexity logarithmically.) In Theorem 19 we make the connection between the two versions of the new definition explicit.

Conditioning on a subset of reasonable points, or equivalently bounding the expectation of the weight function, allows us to base our learnability results on L_1-regularized linear learning. The actual learning rule we get, given in Equation (4.6), is very similar, and even identical, to learning rules suggested by various authors and commonly used in practice as an alternative to Support Vector Machines (Bennett & Campbell, 2000; Roth, 2001; Guigue et al., 2005; Singer, 2000; Tipping, 2001). Here we give a firm theoretical basis to this learning rule, with explicit learning guarantees, and relate it to simple and intuitive properties of the similarity function or kernel used (see the discussion at the end of Section 4).

Structure of this paper: After presenting background on the previous definitions and their relation to kernels in Section 2, we present our new notion of a good similarity function in Section 3. In Section 4 we show that our new broader notions still imply learnability. In Section 5 we give our separation results, showing that our new notion is strictly more general than the notion of a large margin kernel. In Section 6 we show that any large margin kernel is also a good similarity function in our sense, and finally in Section 7 we discuss learning with multiple similarity functions.

2 Background and Notation

We consider a learning problem specified as follows. We are given access to labeled examples (x, ℓ) drawn from some distribution P over $X \times \{-1, 1\}$, where X is an abstract instance space. We will sometimes use D to denote the distribution over x, and for simplicity, we will assume a deterministic target function, so that $(x, \ell) = (x, \ell(x))$. The goal of a learning algorithm is to produce a classification function $g : X \to \{-1, 1\}$ whose error rate $\Pr_{(x,\ell)\sim P}[g(x) \neq \ell]$ is low. We will consider learning algorithms whose only access to points x is through a pairwise similarity function $K(x, x')$ mapping pairs of points to values in the range $[-1, 1]$. Specifically,

Definition 1 *A* similarity function *over X is any pairwise function $K : X \times X \to [-1, 1]$. We say that K is a symmetric similarity function if $K(x, x') = K(x', x)$ for all x, x'.*

Our goal is to describe "goodness" properties that are sufficient for a similarity function to allow one to learn well that ideally are intuitive and subsume the usual notion of good kernel function.

A similarity function K is a valid kernel function if it is positive-semidefinite, i.e. there exists a function ϕ from the instance space X into some (implicit) Hilbert "ϕ-space" such that $K(x, x') = \langle\phi(x), \phi(x')\rangle$. See, e.g., Smola and Schölkopf (2002) for a discussion on conditions for a mapping being a kernel function. Throughout this work, and without loss of generality, we will only consider kernels such that $K(x, x) \leq 1$ for all $x \in \mathcal{X}$ (any kernel K can be converted into this form by, for instance, defining $\tilde{K}(x, x') =$

$K(x,x')/\sqrt{K(x,x)K(x',x')}$). We say that K is (ϵ,γ)-*kernel good* for a given learning problem P if there exists a vector β in the ϕ-space that has error ϵ at margin γ; for simplicity we consider only separators through the origin. Specifically:

Definition 2 *K is an* (ϵ,γ)-good kernel function *if there exists a vector β, $\|\beta\| \leq 1$ such that*

$$\Pr_{(x,\ell)\sim P}[\ell\langle\phi(x),\beta\rangle \geq \gamma] \geq 1-\epsilon.$$

We say that K is γ-*kernel good* if it is (ϵ,γ)-*kernel good* for $\epsilon = 0$; i.e., it has zero error at margin γ.

Given a kernel that is (ϵ,γ)-kernel-good for some learning problem P, a predictor with error rate at most $\epsilon+\epsilon_{\text{acc}}$ can be learned (with high probability) from a sample of[1] $\tilde{\mathcal{O}}\big((\epsilon+\epsilon_{\text{acc}})/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ random examples from P by minimizing the number of margin γ violations on the sample (McAllester, 2003). However, minimizing the number of margin violations on the sample is a difficult optimization problem: it is NP-hard, and even NP-hard to approximate (Arora et al., 1997; Feldman et al., 2006; Guruswami & Raghavendra, 2006). Instead, it is common to minimize the so-called *hinge loss* relative to a margin.

Definition 3 *We say that K is (ϵ,γ)-kernel good* in hinge-loss *if there exists a vector β, $\|\beta\| \leq 1$ such that*

$$\mathbf{E}_{(x,\ell)\sim P}[[1-\ell\langle\beta,\phi(x)\rangle/\gamma]_+] \leq \epsilon,$$

where $[1-z]_+ = \max(1-z,0)$ is the hinge loss.

Given a kernel that is (ϵ,γ)-kernel-good in hinge-loss, a predictor with error rate at most $\epsilon+\epsilon_{\text{acc}}$ can be efficiently learned from a sample of size $\tilde{\mathcal{O}}\big((\epsilon+\epsilon_{\text{acc}})/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ with high probability by minimizing the average hinge loss relative to margin γ on the sample (Bartlett & Mendelson, 2003).

We now present the definition of a good similarity function from (Balcan & Blum, 2006; Srebro, 2007).

Definition 4 (Previous, Margin Violations) *A pairwise function K is an* (ϵ,γ)-**good similarity function** *for a learning problem P if there* exists *a weighting function $w: X \to [0,1]$ such that at least a $1-\epsilon$ probability mass of examples (x,ℓ) satisfy:*

$$\mathbf{E}_{(x',\ell')\sim P}[\ell\ell' w(x')K(x,x')] \quad \geq \quad \gamma. \qquad (2.1)$$

That is, if the underlying distribution is 50/50 positive and negative, this is saying that the average weighted similarity of an example x to random examples x' of its own label should be 2γ larger than the average weighted similarity of x to random examples x' of the other label.

Balcan and Blum (2006) show how a predictor with error rate at most $\epsilon+\epsilon_{\text{acc}}$ can be learned from $\tilde{\mathcal{O}}\big((\epsilon+\epsilon_{\text{acc}})/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ samples using an (ϵ,γ)-good similarity function K: First draw from P an (unlabeled) sample $S = \{x'_1,\ldots,x'_d\}$ of $d = (4/\gamma)^2\ln(4/(\delta\epsilon_{\text{acc}}))$ random "landmarks", and construct the mapping $\phi^S: X \to \mathbb{R}^d$ defined as ${\phi^S}_i(x) = \frac{1}{\sqrt{d}}K(x,x'_i)$, $i \in \{1,\ldots,d\}$. With probability at least $1-\delta$ over the random sample S, the induced distribution $\phi^S(P)$ in R^d has a separator of error at most $\epsilon+\epsilon_{\text{acc}}/2$ at margin at least $\gamma/2$. Now, draw a fresh sample, map it into the transformed space using ϕ^S, and then learn a good linear separator in the transformed space. The total sample complexity is dominated by the $\tilde{\mathcal{O}}\big((\epsilon+\epsilon_{\text{acc}})d/\epsilon_{\text{acc}}^2\big) = \tilde{\mathcal{O}}\big((\epsilon+\epsilon_{\text{acc}})/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ sample complexity of learning in the transformed space, yielding the same overall sample complexity as with an (ϵ,γ)-good kernel function.

The above bounds refer to learning a linear separator by minimizing the error over the training sample. As mentioned earlier, this minimization problem is NP-hard even to approximate. Again, we can instead consider the hinge-loss rather than the number of margin violations. Balcan and Blum (2006) and Srebro (2007) therefore provide the following hinge-loss version of their definition:

Definition 5 (Previous, Hinge Loss) *A similarity function K is an* (ϵ,γ)-**good similarity function in hinge loss** *for a learning problem P if there exists a weighting function $w(x')$ $\in [0,1]$ for all $x' \in X$ such that*

$$\mathbf{E}_{(x,\ell)\sim P}\Big[\,[1-\ell g(x)/\gamma]_+\,\Big] \leq \epsilon, \qquad (2.2)$$

where $g(x) = \mathbf{E}_{(x',\ell')\sim P}[\ell' w(x')K(x,x')]$ is the similarity-based prediction made using $w()$, and recall that $[1-z]_+ = \max(0,1-z)$ is the hinge-loss.

The same algorithm as above, but now using SVM to minimize hinge-loss in the transformed space, allows one to efficiently use a similarity function satisfying this definition to find a predictor of error $\epsilon+\epsilon_{\text{acc}}$ using $\tilde{\mathcal{O}}\big((\epsilon+\epsilon_{\text{acc}})/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ examples.

3 New Notions of Good Similarity Functions

In this section we provide new notions of good similarity functions generalizing Definitions 4 and 5 that we prove have a number of important advantages.

In the definitions of Balcan and Blum (2006), a weight $w(x') \in [0,1]$ was used in defining the quantity of interest $\mathbf{E}_{(x',\ell')\sim P}[\ell' w(x')K(x,x')]$. Here, it will instead be more convenient to think of w as the expected value of an indicator random variable $R(x) \in \{0,1\}$ where we will view the (probabilistic) set $\{x : R(x) = 1\}$ as a set of "reasonable points". Formally, we will then be sampling from the joint distribution $P(x,\ell(x),R(x)) = P(x,\ell(x))P(R(x)|x)$ but we will sometimes omit the explicit dependence on R when it is clear from context. Our new definition is now as follows.

Definition 6 (Main, Margin Violations) *A similarity function K is an* (ϵ,γ,τ)-**good similarity function** *for a learning problem P if there* exists *a (random) indicator function $R(x)$ defining a (probabilistic) set of "reasonable points" such that the following conditions hold:*

1. *A $1-\epsilon$ probability mass of examples (x,ℓ) satisfy*

$$\mathbf{E}_{(x',\ell')\sim P}[\ell\ell' K(x,x') \mid R(x')] \geq \gamma \qquad (3.1)$$

2. $\Pr_{x'}[R(x')] \geq \tau$.

[1] The $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic factors in the arguments and in the failure probability.

If the reasonable set R is $50/50$ positive and negative (i.e., $\Pr_{x'}[\ell(x') = 1|R(x')] = 1/2$), we can interpret the condition as stating that most examples x are on average 2γ more similar to random reasonable examples x' of their own label than to random reasonable examples x' of the other label. The second condition is that at least a τ fraction of the points should be reasonable.

We also consider a hinge-loss version of the definition:

Definition 7 (Main, Hinge Loss) *A similarity function K is an (ϵ, γ, τ)-**good similarity function in hinge loss** for a learning problem P if there* exists *a (probabilistic) set R of "reasonable points" such that the following conditions hold:*

1. *We have*

$$\mathbf{E}_{(x,\ell)\sim P}\Big[\,[1 - \ell g(x)/\gamma]_+\,\Big] \leq \epsilon, \qquad (3.2)$$

   where $g(x) = \mathbf{E}_{(x',\ell',R(x'))}[\ell' K(x, x') \mid R(x')]$.
2. $\Pr_{x'}[R(x')] \geq \tau$.

It is not hard to see that an (ϵ, γ)-good similarity function under Definitions 4 and 5 is also an $(\epsilon, \gamma, \gamma)$-good similarity function under Definitions 6 and 7, respectively. In the reverse direction, an (ϵ, γ, τ)-good similarity function under Definitions 6 and 7 is an $(\epsilon, \gamma\tau)$-good similarity function under Definitions 4 and 5 (respectively). For formal proofs, see Theorems 23 and 24 in Appendix A.

As we will see, under both old and new definitions, the number of labeled samples required for learning grows as $1/\gamma^2$. The key distinction between them is that we introduce a new parameter, τ, that primarily affects the number of *unlabeled* examples required. This decoupling of the number of labeled and unlabeled examples enables us to handle a wider variety of situations with an improved labeled sample complexity. In particular, in translating from a kernel to a similarity function, we will find that much of the loss can now be placed into the τ parameter.

In the following we prove three types of results about this new notion of similarity. The first is that similarity functions satisfying these conditions are sufficient for learning (in polynomial time in the case of Definition 7), with a sample size of $O(\frac{1}{\gamma^2}\ln(\frac{1}{\gamma\tau}))$ labeled examples and $O(\frac{1}{\tau\gamma^2})$ unlabeled examples. This is particularly useful in settings where unlabeled data is plentiful and cheap—such settings are increasingly common in learning applications (Mitchell, 2006; Chapelle et al., 2006)—or for distribution-specific learning where unlabeled data may be viewed as free.

The second main theorem we prove is that *any* class C, over a sufficiently unconcentrated distribution on examples, has a $(0, 1, 1/(2|C|))$-good similarity function (under either definition 6 or 7), whereas there exist classes C that have no $(0.5, 8/\sqrt{|C|})$-good kernel functions in hinge loss. This provides a clear separation between the similarity and kernel notions in terms of the parameters controlling labeled sample complexity. The final main theorem we prove is that any large-margin kernel function also satisfies our similarity definitions, with substantially less loss in the parameters controlling labeled sample complexity compared to the definition of (Balcan & Blum, 2006). For example, if K is a $(0, \gamma)$-good kernel, then it is an $(\epsilon', \epsilon'\gamma^2)$-good similarity function under Definitions 4 and 5, and this is tight (Srebro, 2007), resulting in a sample complexity of $\tilde{\mathcal{O}}\big(1/(\gamma^4\epsilon^3)\big)$ to achieve error ϵ. However, we can show K is an $(\epsilon', \gamma^2, \epsilon')$-good similarity function under the new definition,[2] resulting in a sample complexity of only $\tilde{\mathcal{O}}\big(1/(\gamma^4\epsilon)\big)$.

4 Good Similarity Functions Allow Learning

The basic approach proposed for learning using a similarity function is similar to that of Balcan and Blum (2006). First, a feature space is constructed, consisting of similarities to randomly chosen landmarks. Then, a linear predictor is sought in this feature space. However, under the previous definitions, we were guaranteed large L_2-margin in this feature space, whereas under the new definitions we are guaranteed large L_1-margin in the feature space.

After recalling the notion of an L_1-margin and its associated learning guarantee, we first establish that, for an (ϵ, γ, τ)-good similarity function, the feature map constructed using $\tilde{\mathcal{O}}\big(1/(\tau\gamma^2)\big)$ landmarks indeed has (with high probability) a large L_1-margin separator. Using this result, we then obtain a learning guarantee by following the strategy outlined above.

In speaking of L_1-margin γ, we refer to separation with a margin γ by a unit-L_1-norm linear separator, in a unit-L_∞-bounded feature space. Formally, let $\phi : x \mapsto \phi(x)$, $\phi(x) \in \mathbb{R}^d$, with $\|\phi(x)\|_\infty \leq 1$ be a mapping of the data to a d-dimensional feature space. We say that a linear predictor $\alpha \in \mathbb{R}^d$, achieves error ϵ relative to L_1-margin γ if $\Pr_{x,\ell(x)}(\ell(x)\langle\alpha, \phi(x)\rangle \geq \gamma) \geq 1 - \epsilon$ (this is the standard margin constraint) and $\|\alpha\|_1 = 1$.

Given a d-dimensional feature map under which there exists some (unknown) zero-error linear separator with L_1-margin γ, we can efficiently learn a predictor with error at most ϵ_{acc} using $O\left(\frac{\log d}{\epsilon_{\text{acc}}\gamma^2}\right)$ examples (with high probability). This can be done using the Winnow algorithm with a standard online-to-batch conversion (Littlestone, 1989). If we can only guarantee the existence of a separator with error $\epsilon > 0$ relative to L_1-margin γ, then a predictor with error $\epsilon + \epsilon_{\text{acc}}$ can be theoretically learned (with high probability) from a sample of $\tilde{\mathcal{O}}\big((\log d)/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ examples by minimizing the number of L_1-margin γ violations on the sample (Zhang, 2002).

We are now ready to state the main result enabling learning using good similarity functions:

Theorem 8 *Let K be an (ϵ, γ, τ)-good similarity function for a learning problem P. Let $S = \{x'_1, x'_2, \ldots, x'_d\}$ be a (potentially unlabeled) sample of*

$$d = \frac{2}{\tau}\left(\log(2/\delta) + 8\frac{\log(2/\delta)}{\gamma^2}\right)$$

landmarks drawn from P. Consider the mapping $\phi^S : X \to \mathbb{R}^d$ defined as follows: $\phi^S{}_i(x) = K(x, x'_i)$, $i \in \{1, \ldots, d\}$. Then, with probability at least $1 - \delta$ over the random sample S, the induced distribution $\phi^S(P)$ in R^d has a separator of error at most $\epsilon + \delta$ relative to L_1 margin at least $\gamma/2$.

[2] Formally, the translation produces an $(\epsilon', \gamma^2/c, \epsilon' c)$-good similarity function for some $c \leq 1$. However, smaller values of c only improve the bounds.

Proof: First, note that since $|K(x,x)| \leq 1$ for all x, we have $\left\|\phi^S(x)\right\|_\infty \leq 1$.

Consider the linear separator $\alpha \in \mathbb{R}^d$, given by $\alpha_i = \ell(x'_i)R(x'_i)/d_1$ where $d_1 = \sum_i R(x'_i)$ is the number of landmarks with $R(x') = 1$. This normalization ensures $\|\alpha\|_1 = 1$. Note that we take $R(x'_i)$ to be drawn jointly with x'_i. If it is random, than it is randomly instantiated to either zero or one.

We have, for any $x, \ell(x)$:

$$\ell(x)\langle \alpha, \phi^S(x)\rangle = \frac{\sum_{i=1}^d \ell(x)\ell(x'_i)R(x'_i)K(x,x'_i)}{d_1} \quad (4.1)$$

This is an empirical average of d_1 terms

$$-1 \leq \ell(x)\ell(x')K(x,x') \leq 1$$

for which $R(x') = 1$. For any x we can apply Hoeffding's inequality, and obtain that with probability at least $1 - \delta^2/2$ over the choice of S, we have:

$$\ell(x)\langle \alpha, \phi^S(x)\rangle \geq \mathbf{E}_{x'}[K(x,x')\ell(x')\ell(x)|R(x')] - \sqrt{\frac{2\log(\frac{2}{\delta^2})}{d_1}} \quad (4.2)$$

Since the above holds for any x with probability at least $1-\delta^2/2$ over S, it also holds with probability at least $1-\delta^2/2$ over the choice of x and S. We can write this as:

$$\mathbf{E}_{S\sim P^d}\left[\Pr_{x\sim P}(\text{ violation })\right] \leq \delta^2/2 \quad (4.3)$$

where "violation" refers to violating (4.2). Applying Markov's inequality we get that with probability at least $1 - \delta/2$ over the choice of S, at most δ fraction of points violate (4.2). Recalling Definition 6, at most an additional ϵ fraction of the points violate (3.1). But for the remaining $1 - \epsilon - \delta$ fraction of the points, for which both (4.2) and (3.1) hold, we have:

$$\ell(x)\langle \alpha, \phi^S(x)\rangle \geq \gamma - \sqrt{\frac{2\log(\frac{1}{\delta^2})}{d_1}}. \quad (4.4)$$

To bound the second term we need an upper bound on d_1, the number of reasonable landmarks. The probability of each of the d landmarks being reasonable is at least τ and so the number of reasonable landmarks follows a Binomial distribution, ensuring $d_1 \geq 8\log(1/\delta)/\gamma^2$ with probability at least $1 - \delta/2$. When this happens, we have $\sqrt{\frac{2\log(\frac{1}{\delta^2})}{d_1}} \leq \gamma/2$. We get then, that with probability at least $1 - \delta$, for at least $1 - \epsilon - \delta$ of the points:

$$\ell(x)\langle \alpha, \phi^S(x)\rangle \geq \gamma/2. \quad (4.5)$$

■

For the realizable ($\epsilon = 0$) case, we obtain:

Corollary 9 *If K is an $(0,\gamma,\tau)$-good similarity function then with high probability we can efficiently find a predictor with error at most ϵ_{acc} from an unlabeled sample of size $d_u = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2\tau}\right)$ and from a labeled sample of size $d_l = \tilde{\mathcal{O}}\left(\frac{\log d_u}{\gamma^2\epsilon_{acc}}\right)$.*

Proof: We have proved in Theorem 8 that if K is$(0,\gamma,\tau)$-good similarity function, then with high probability there exists a low-error (at most δ) large-margin (at least $\frac{\gamma}{2}$) separator in the transformed space under mapping ϕ^S. Thus, all we need now to learn well is to draw a new fresh sample $\tilde{S}$, map it into the transformed space using ϕ^S, and then apply a good algorithm for learning linear separators in the new space that produces a hypothesis of error at most ϵ_{acc} with probability at least $1 - \delta$. In particular, remember that the vector α has error at most δ at L_1 margin $\gamma/2$ over $\phi^S(P)$, where the mapping ϕ^S produces examples of L_∞ norm at most 1. In order to enjoy the better learning guarantees of the separable case, we will set δ small enough so that no bad points appear in the sample. The Corollary now follows from the L_1-margin learning guarantee in the separable case, discussed earlier in the Section. ■

For the general ($\epsilon > 0$) case, Theorem 8 implies that by following our two-stage approach, first using $d_u = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2\tau}\right)$ unlabeled examples as landmarks in order to construct $\phi^S(\cdot)$, and then using a fresh sample of size $d_l = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2\epsilon_{\text{acc}}^2}\ln d_u\right)$ to learn a low-error L_1-margin γ separator in $\phi^S(\cdot)$, we have:

Corollary 10 *If K is a (ϵ,γ,τ)-good similarity function then by minimizing L_1 margin violations we can find a predictor with error at most ϵ_{acc} from an unlabeled sample of size $d_u = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2\tau}\right)$ and from a labeled sample of size $d_l = \tilde{\mathcal{O}}\left(\frac{\log d_u}{\gamma^2\epsilon_{acc}^2}\right)$.*

The procedure described above, although well defined, involves a difficult optimization problem: minimizing the number of L_1-margin violations. In order to obtain a computationally tractable procedure, we consider the hinge-loss instead of the margin error. In a feature space with $\|\phi(x)\|_\infty \leq 1$ as above, we say that a unit-L_1-norm predictor α, $\|\alpha\|_1 = 1$, has expected hinge-loss $\mathbf{E}\left[[1 - \ell(x)\langle\alpha,\phi(x)\rangle/\gamma]_+\right]$ relative to L_1-margin γ. Now, if we know there is some (unknown) predictor with hinge-loss ϵ relative L_1-margin γ, than a predictor with error $\epsilon + \epsilon_{\text{acc}}$ can be learned (with high probability) from a sample of $\tilde{\mathcal{O}}\left(\log d/(\gamma^2\epsilon_{\text{acc}}^2)\right)$ examples by minimizing the empirical average hinge-loss relative to L_1-margin γ on the sample (Zhang, 2002).

Before proceeding to discussing the optimization problem of minimizing the average hinge-loss relative to a fixed L_1-margin, let us establish the analogue of Theorem 8 for the hinge-loss:

Theorem 11 *Let K be an (ϵ,γ,τ)-good similarity function in hinge-loss for a learning problem P. For any $\epsilon_1 > 0$ and $0 < \lambda < \gamma\epsilon_1/4$ let $S = \{x'_1, x'_2, \ldots, x'_d\}$ be a sample of size $d = \frac{2}{\tau}\left(\log(2/\delta) + 16\log(2/\delta)/(\epsilon_1\gamma)^2\right)$ drawn from P. With probability at least $1 - \delta$ over the random sample S, the induced distribution $\phi^S(P)$ in R^d, for ϕ^S as defined in Theorem 8, has a separator achieving hinge-loss at most $\epsilon + \epsilon_1$ at margin γ.*

Proof: We use the same construction as in Theorem 8. ■

Corollary 12 *K is an (ϵ, γ, τ)-good similarity function in hinge loss then we can efficiently find a predictor with error at most $\epsilon + \epsilon_{acc}$ from an unlabeled sample of size $d_u = \tilde{\mathcal{O}}\left(\frac{1}{\gamma^2 \epsilon_{acc}^2 \tau}\right)$ and from a labeled sample of size $d_l = \tilde{\mathcal{O}}\left(\frac{\log d_u}{\gamma^2 \epsilon_{acc}^2}\right)$.*

For the hinge-loss, our two stage procedure boils down to solving the following optimization problem w.r.t. α:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{d_l} \left[1 - \sum_{j=1}^{d_u} \alpha_j \ell(x_i) K(x_i, x'_j) \right]_+ \\ \text{s.t.} \quad & \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma \end{aligned} \tag{4.6}$$

This is a linear program and can thus be solved in polynomial time, establishing the efficiency in Corollary 12.

An optimization problem similar to (4.6), though usually with the same set of points used both as landmarks and as training examples, is actually fairly commonly used as a learning rule in practice (Bennett & Campbell, 2000; Roth, 2001; Guigue et al., 2005). Such a learning rule is typically discussed as an alternative to SVMs. In fact, Tipping (2001) suggest the Relevance Vector Machine (RVM) as a Bayesian alternative to SVMs. The MAP estimate of the RVM is given by an optimization problem similar to (4.6), though with a loss function different from the hinge loss (the hinge-loss cannot be obtained as a log-likelihood). Similarly, Singer (2000) suggests Norm-Penalized Leveraging Procedures as a boosting-like approach that mimics SVMs. Again, although the specific loss functions studied by Singer are different from the hinge-loss, the method (with a norm exponent of 1, as in Singer's experiments) otherwise corresponds to a coordinate-descent minimization of (4.6). In both cases, no learning guarantees are provided.

The motivation for using (4.6) as an alternative to SVMs is usually that the L_1-regularization on α leads to sparsity, and hence to "few support vectors" (although Vincent and Bengio (2002), who also discuss (4.6), argue for more direct ways of obtaining such sparsity), and also that the linear program (4.6) might be easier to solve than the SVM quadratic program. However, we are not aware of a previous discussion on how learning using (4.6) relates to learning using a SVM, or on learning guarantees using (4.6) in terms of properties of the similarity function K. Guarantees solely in terms of the feature space in which we seek low L_1-margin (ϕ^S in our notation) are problematic, as this feature space is generated randomly from data.

In fact, in order to enjoy the SVM guarantees while using L_1 regularization to obtain sparsity, some authors suggest regularizing both the L_1 norm $\|\alpha\|_1$ of the coefficient vector α (as in (4.6)), and the norm $\|\beta\|$ of the corresponding predictor $\beta = \sum_j \alpha_j \phi(x'_j)$ in the Hilbert space implied by K, where $K(x, x') = \langle \phi(x), \phi(x') \rangle$, as when using a SVM with K as a kernel (Osuna & Girosi, 1999; Gunn & Kandola, 2002).

Here, we provide a natural condition on the similarity function K (Definition 7), that justifies the learning rule (4.6). Furthermore, we show (in Section 6) than any similarity function that is good as a kernel, and can ensure SVM learning, is also good as a similarity function and can thus also ensure learning using the learning rule (4.6) (though possibly with some deterioration of the learning guarantees). These arguments can be used to justify (4.6) as an alternative to SVMs.

Before concluding this discussion, we would like to mention that Girosi (1998) previously established a rather different connection between regularizing the L_1 norm $\|\alpha\|_1$ and regularizing the norm of the corresponding predictor β in the implied Hilbert space. Girosi considered a hard-margin SVR (Support Vector Regression Machine, i.e. requiring each prediction to be within $(\ell(x) - \epsilon, \ell(x) + \epsilon)$), in the noiseless case where the mapping $x \mapsto \ell(x)$ is in the Hilbert space. In this setting, Girosi showed that a hard-margin SVR is equivalent to minimizing the distance *in the implied Hilbert space* between the correct mapping $x \mapsto \ell(x)$ and the predictions $x \mapsto \sum_j \alpha_j K(x, x'_j)$, with an L_1 regularization term $\|\alpha\|_1$. However, this distance between prediction functions is very different than the objective in (4.6), and again refers back to the implied feature space which we are trying to avoid.

5 Separation Results

In this Section, we show an example of a finite concept class for which no kernel yields good learning guarantees when used as a kernel, but for which there does exist a good similarity function yielding the optimal sample complexity. That is, we show that some concept classes cannot be reasonably represented by kernels, but can be reasonably represented by similarity functions.

Specifically, we consider a class C of n pairwise uncorrelated functions. This is a finite class of cardinality $|C| = n$, and so if the target belongs to C then $O(\frac{1}{\epsilon} \log n)$ samples are enough for learning a predictor with error ϵ.

Indeed, we show here that for *any* concept class C, so long as the distribution D is sufficiently unconcentrated, there exists a similarity function that is $(0, 1, \frac{1}{2|C|})$-good under our definition for every $f \in C$. This yields a (labeled) sample complexity $O(\frac{1}{\epsilon} \log |C|)$ to achieve error ϵ, matching the ideal sample complexity. In other words, for distribution-specific learning (where unlabeled data may be viewed as free) and finite classes, there is no *intrinsic* loss in sample-complexity incurred by choosing to learn via similarity functions. In fact, we also extend this result to classes of bounded VC-dimension rather than bounded cardinality.

In contrast, we show that if C is a class of n functions that are pairwise uncorrelated with respect to distribution D, then *no* kernel is (ϵ, γ)-good in hinge-loss for all $f \in C$ even for $\epsilon = 0.5$ and $\gamma = 8/\sqrt{n}$. This extends work of (Ben-David et al., 2003; Forster & Simon, 2006) who give hardness results with comparable margin bounds, but at a much lower error rate. Thus, this shows there *is* an intrinsic loss incurred by using kernels together with margin bounds, since this results in a sample complexity bound of at least $\Omega(|C|)$, rather than the ideal $O(\log |C|)$.

We thus demonstrate a gap between the kind of prior knowledge can be represented with kernels as opposed to general similarity functions and demonstrate that similarity functions are strictly more expressive (up to the degradation in parameters discussed earlier).

Definition 13 *We say that a distribution D over X is α-*

unconcentrated *if the probability mass on any given $x \in X$ is at most α.*

Theorem 14 *For any class finite class of functions C and for any $1/|C|$-unconcentrated distribution D over the instance space X, there exists a similarity function K that is a $(0, 1, \frac{1}{2|C|})$-good similarity function for all $f \in C$.*

Proof: Let $C = \{f_1, \ldots, f_n\}$. Now, let us partition X into n regions R_i of at least $1/(2n)$ probability mass each, which we can do since D is $1/n$-unconcentrated. Finally, define $K(x, x')$ for x' in R_i to be $f_i(x)f_i(x')$. We claim that for this similarity function, R_i is a set of "reasonable points" establishing margin $\gamma = 1$ for target f_i. Specifically,

$$\begin{aligned}\mathbf{E}[K(x,x')f_i(x)f_i(x')|x' \in R_i] &= \mathbf{E}[f_i(x)f_i(x')f_i(x)f_i(x')] \\ &= 1.\end{aligned}$$

Since $\Pr(R_i) \geq \frac{1}{2n}$, this implies that under distribution D, K is a $(0, 1, \frac{1}{2n})$-good similarity function for all $f_i \in C$. ■

Note 1: We can extend this argument to any class C of small VC dimension. In particular, for any distribution D, the class C has an ϵ-cover C_ϵ of size $(1/\epsilon)^{O(d/\epsilon)}$, where d is the VC-dimension of C (Benedek & Itai, 1988). By Theorem 14, we can have a $(0, 1, 1/|C_\epsilon|)$-good similarity function for the cover C_ϵ, which in turn implies an $(\epsilon, 1, 1/|C_\epsilon|)$-good similarity function for the original set (even in hinge loss since $\gamma = 1$). Plugging in our bound on $|C_\epsilon|$, we get an $(\epsilon, 1, \epsilon^{O(d/\epsilon)})$-good similarity function for C. Thus, the labeled sample complexity we get for learning with similarity functions is only $\mathcal{O}((d/\epsilon)\log(1/\epsilon))$, and again there is no *intrinsic* loss in sample complexity bounds due to learning with similarity functions.

Note 2: The need for the underlying distribution to be unconcentrated stems from our use of this distribution for both labeled and unlabeled data. We could further extend our definition of "good similarity function" to allow for the unlabeled points x' to come from some other distribution D' *given apriori*, such as the uniform distribution over the instance space X. Now, the expectation over x' and the probability mass of R would both be with respect to D', and the generic learning algorithm would draw points x'_i from D' rather than D. In this case, we would only need D' to be unconcentrated, rather than D.

We now prove our lower bound for margin-based learning with kernels.

Theorem 15 *Let C be a class of n pairwise uncorrelated functions over distribution D. Then, there is no kernel that for all $f \in C$ is (ϵ, γ)-good in hinge-loss even for $\epsilon = 0.5$ and $\gamma = 8/\sqrt{n}$.*

Proof: Let $C = \{f_1, \ldots, f_n\}$. We begin with the basic fourier setup (Linial et al., 1989; Mansour, 1994). Given two functions f and g, define $\langle f, g\rangle = \mathbf{E}_x[f(x)g(x)]$ to be their correlation with respect to distribution D. (This is their inner-product if we view f as a vector whose jth coordinate is $f(x_j)[D(x_j)]^{1/2}$). Because the functions $f_i \in C$ are pairwise uncorrelated, we have $\langle f_i, f_j\rangle = 0$ for all $i \neq j$, and because the f_i are boolean functions we have $\langle f_i, f_i\rangle = 1$ for all i. Thus they form at least part of an orthonormal basis, and for any hypothesis h (i.e. any mapping $X \to \{\pm 1\}$) we have

$$\sum_{f_i \in C} \langle h, f_i\rangle^2 \leq 1.$$

So, this implies

$$\sum_{f_i \in C} |\langle h, f_i\rangle| \leq \sqrt{n}.$$

or equivalently

$$\mathbf{E}_{f_i \in C}|\langle h, f_i\rangle| \leq 1/\sqrt{n}. \quad (5.1)$$

In other words, for any hypothesis h, if we pick the target at random from C, the expected magnitude of the correlation between h and the target is at most $1/\sqrt{n}$.

We now consider the implications of having a good kernel. Suppose for contradiction that there exists a kernel K that is $(0.5, \gamma)$-good in hinge loss for every $f_i \in C$. What we will show is this implies that for any $f_i \in C$, the expected value of $|\langle h, f_i\rangle|$ for a *random* linear separator h in the ϕ-space is greater than $\gamma/8$. If we can prove this, then we are done because this implies there must *exist* an h that has $\mathbf{E}_{f_i \in C}|\langle h, f\rangle| > \gamma/8$, which contradicts equation (5.1) for $\gamma = 8/\sqrt{n}$.

So, we just have to prove the statement about random linear separators. Let w^* denote the vector in the ϕ-space that has hinge-loss at most 0.5 at margin γ for target function f_i. For any example x, define γ_x to be the margin of $\phi(x)$ with respect to w^*, and define $\alpha_x = \sin^{-1}(\gamma_x)$ to be the angular margin of $\phi(x)$ with respect to w^*.[3] Now, consider choosing a random vector h in the ϕ-space, where we associate $h(x) = \text{sign}(\text{h} \cdot \phi(\text{x}))$. Since we only care about the absolute value $|\langle h, f_i\rangle|$, and since $\langle -h, f_i\rangle = -\langle h, f_i\rangle$, it suffices to show that $\mathbf{E}_h[\langle h, f_i\rangle \mid h \cdot w^* \geq 0] > \gamma/8$. We do this as follows.

First, for any example x, we claim that:

$$\Pr_h[(h(x) \neq f_i(x)|h \cdot w^* \geq 0] = 1/2 - \alpha_x/\pi. \quad (5.2)$$

This is because we look at the 2-dimensional plane defined by $\phi(x)$ and w^*, and consider the half-circle of $\|h\| = 1$ such that $h \cdot w^* \geq 0$, then (5.2) is the portion of the half-circle that labels $\phi(x)$ incorrectly. Thus, we have:

$$\mathbf{E}_h[err(h)|h \cdot w^* \geq 0] \;=\; \mathbf{E}_x[1/2 - \alpha_x/\pi],$$

and so, using $\langle h, f_i\rangle = 1 - 2\,err(h)$, we have:

$$\mathbf{E}_h[\langle h, f_i\rangle \mid h \cdot w^* \geq 0] \;=\; 2\mathbf{E}_x[\alpha_x]/\pi.$$

Finally, we just need to relate angular margin and hinge loss: if L_x is the hinge-loss of $\phi(x)$, then a crude bound on α_x is

$$\alpha_x \geq \gamma(1 - (\pi/2)L_x).$$

[3] So, α_x is a bit larger in magnitude than γ_x. This works in our favor when the margin is positive, and we just need to be careful when the margin in negative.

Since we assumed that $\mathbf{E}_x[L_x] \leq 0.5$, we have:

$$\mathbf{E}_x[\alpha_x] \geq \gamma(1 - \pi/4).$$

Putting this together we get expected magnitude of correlation of a random halfspace is at least $2\gamma(1 - \pi/4)/\pi > \gamma/8$ as desired, proving the theorem. ■

An example of a class C satisfying the above conditions is the class of parity functions over $\{0,1\}^{\lg n}$, which are pairwise uncorrelated with respect to the uniform distribution. Note that the uniform distribution is $1/|C|$-unconcentrated, and thus there *is* a good similarity function. (In particular, one could use $K(x_i, x_j) = f_j(x_i)f_j(x_j)$, where f_j is the parity function associated with indicator vector x_j.)

We can extend Theorem 15 to classes of large Statistical Query dimension as well. In particular, the SQ-dimension of a class C with respect to distribution D is the size d of the largest set of functions $\{f_1, f_2, \ldots, f_d\} \subseteq C$ such that $|\langle f_i, f_j\rangle| \leq 1/d^3$ for all $i \neq j$ (Blum et al., 1994). In this case, we just need to adjust the Fourier analysis part of the argument to handle the fact that the functions may not be completely uncorrelated.

Theorem 16 *Let C be a class of functions of SQ-dimension d with respect to distribution D. Then, there is no kernel that for all $f \in C$ is (ϵ, γ)-good in hinge-loss even for $\epsilon = 0.5$ and $\gamma = 16/\sqrt{d}$.*

Proof: Let $f_1, \ldots, f_d$ be d functions in C such that $|\langle f_i, f_j\rangle| \leq 1/d^3$ for all $i \neq j$. We can define an orthogonal set of functions $f'_1, f'_2, \ldots, f'_d$ as follows: let $f'_1 = f_1$, $f'_2 = f_2 - f_1\langle f_2, f_1\rangle$, and in general let f'_i be the portion of f_i orthogonal to the space spanned by $f_1, \ldots, f_{i-1}$. (That is, $f'_i = f_i - \mathrm{proj}(f_i, \mathrm{span}(f_1, \ldots, f_{i-1}))$, where "proj" is orthogonal projection.) Since the f'_i are orthogonal and have length at most 1, for any boolean function h we have $\sum_i \langle h, f'_i\rangle^2 \leq 1$ and therefore $\mathbf{E}_i|\langle h, f'_i\rangle| \leq 1/\sqrt{d}$. Finally, since $\langle f_i, f_j\rangle \leq 1/d^3$ for all $i \neq j$, one can show this implies that $|f_i - f'_i| \leq 1/d$ for all i. So, $\mathbf{E}_i|\langle h, f_i\rangle| \leq 1/\sqrt{d} + 1/d \leq 2/\sqrt{d}$. The rest of the argument in the proof of Theorem 15 now applies with $\gamma = 16/\sqrt{d}$. ■

For example, the class of size-n decision trees over $\{0,1\}^n$ has $n^{\Omega(\log n)}$ pairwise uncorrelated functions over the uniform distribution (in particular, any parity of $\log n$ variables can be written as an n-node decision tree). So, this means we cannot have a kernel with margin $1/poly(n)$ for all size-n decision trees over $\{0,1\}^n$. However, we *can* have a similarity function with margin 1, though the τ parameter (which controls running time) will be exponentially small.

6 Relation between kernels and similarity functions

As is shown in the Appendix (Theorem 25), if a similarity function K is indeed a kernel, and it is (ϵ, γ, τ)-good as a similarity function (possibly in hinge-loss), than it is also (ϵ, γ)-good as a kernel (respectively, in hinge loss). That is, although the notion of a good similarity function is more widely applicable, for those similarity functions that are positive semidefinite, a good similarity function is also a good kernel. We now show the converse: if a kernel function is good in the kernel sense, it is also good in the similarity sense, though with some degradation of the margin. This degradation is much smaller than the one incurred previously by Balcan and Blum (2006) and Srebro (2007). Specifically, we can show that if K is a $(0, \gamma)$-good kernel, then K is $(\epsilon, \gamma^2, \epsilon)$-good similarity function for any ϵ (formally, it is $(\epsilon, \gamma^2/c, \epsilon c)$-good for some $c \leq 1$).

To prove this relationship, we introduce an intermediate notion of a good similarity function.

Definition 17 (Intermediate, Margin Violations) *A similarity function K is a* **strongly (ϵ, γ, M)-good similarity function** *for a learning problem P if there* exists *a bounded weighting function w over X, $w(x') \in [0, M]$ for all $x' \in X$, $\mathbf{E}[w] \leq 1$ such that at least a $1 - \epsilon$ probability mass of examples x satisfy:*

$$\mathbf{E}_{x' \sim P}[\ell(x)\ell(x')w(x')K(x, x')] \quad \geq \quad \gamma. \qquad (6.1)$$

Definition 18 (Intermediate, Hinge Loss) *A similarity function K is a* **strongly (ϵ, γ, M)-good similarity function in hinge loss** *for a learning problem P if there exists a weighting function $w(x') \in [0, M]$ for all $x' \in X$, $\mathbf{E}[w] \leq 1$ such that*

$$\mathbf{E}_x\Big[\,[1 - \ell(x)g(x)/\gamma]_+\,\Big] \leq \epsilon, \qquad (6.2)$$

where $g(x) = \mathbf{E}_{x' \sim P}[\ell(x')w(x')K(x, x')]$ is the similarity-based prediction made using $w(\cdot)$.

These intermediate definitions are closely related to our main similarity function definitions: in particular, if K is a strongly (ϵ, γ, M)-good similarity function for a learning problem P, then it is also an $(\epsilon, \gamma/c, c/M)$-good similarity function for some $\gamma \leq c \leq 1$.

Theorem 19 *If K is a strongly (ϵ, γ, M)-good similarity function for a learning problem P, then there exists $\gamma \leq c \leq 1$ such that K is a $(\epsilon, \gamma/c, c/M)$-good similarity function for P. If K is a strongly (ϵ, γ, M)-good similarity function in hinge loss for P, then there exists $\gamma \leq c \leq 1$ such that K is a $(\epsilon, \gamma/c, c/M)$-good similarity function for P.*

Note that since our guarantees for (ϵ, γ, τ)-good similarity functions depend on τ only through $\gamma^2\tau$, a decrease in τ and a proportional increase in γ (as when $c < 1$ in Theorem 19) only improves the guarantees. However, allowing flexibility in this tradeoff will make the kernel-to-similarity function translation much easier.

Proof: ***(of Theorem 19)*** First, divide w by M to scale its range to $[0, 1]$, so $\mathbf{E}[w] = c/M$ for some $c \leq 1$ and the margin is now γ/M. Define random indicator $R(x')$ to equal 1 with probability $w(x')$ and 0 with probability $1 - w(x')$, so we have

$$\tau = \Pr_{x'}[R(x') = 1] = \mathbf{E}[w] = c/M,$$

and we can rewrite (6.1) as

$$\mathbf{E}_{x' \sim P,R}[\ell(x)\ell(x')R(x')K(x, x')] \quad \geq \quad \gamma/M. \quad (6.3)$$

Finally, divide both sides of (6.3) by $\tau = c/M$, producing the conditional $\mathbf{E}_{x'}[\ell(x)\ell(x')K(x, x') \mid R(x')]$ on the LHS

and a margin of γ/c on the RHS. The case of hinge-loss is identical. ■

We will now establish that a similarity function K that is good as a kernel, is also good as a similarity function in this intermediate sense, and hence, by Theorem 19, also in our original sense. We begin by considering goodness in hinge-loss, and will return to margin violations at the end of the Section.

Theorem 20 *If K is (ϵ_0, γ)-good kernel in hinge loss for learning problem (with deterministic labels), then it is also a strongly $(\epsilon_0 + \epsilon_1, \frac{\gamma^2}{1+\epsilon_0/2\epsilon_1}, \frac{1}{2\epsilon_1+\epsilon_0})$-good similarity in hinge loss for the learning problem, for any $\epsilon_1 > 0$.*

Proof: We initially only consider finite discrete distributions, where:

$$\Pr(\, x_i, y_i \,) = p_i \tag{6.4}$$

for $i = 1 \ldots n$, with $\sum_{i=1}^{n} p_i = 1$ and $x_i \neq x_j$ for $i \neq j$.

Let K be any kernel function that is (ϵ_0, γ)-kernel good in hinge loss. Let ϕ be the implied feature mapping and denote $\phi_i = \phi(x_i)$. Consider the following weighted-SVM quadratic optimization problem with regularization parameter C:

$$\text{minimize } \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} p_i[1 - y_i\langle\beta, \phi_i\rangle]_+ \tag{6.5}$$

The dual of this problem, with dual variables α_i, is:

$$\begin{aligned} \text{maximize } & \sum_i \alpha_i - \frac{1}{2}\sum_{ij} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ \text{subject to } & \; 0 \leq \alpha_i \leq C p_i \end{aligned} \tag{6.6}$$

There is no duality gap, and furthermore the primal optimum β^* can be expressed in terms of the dual optimum α^*: $\beta^* = \sum_i \alpha_i^* y_i \phi_i$.

Since K is (ϵ_0, γ)-kernel-good in hinge-loss, there exists a predictor $\|\beta_0\| = 1$ with average-hinge loss ϵ_0 relative to margin γ. The primal optimum β^* of (6.5), being the optimum solution, then satisfies:

$$\begin{aligned} &\frac{1}{2}\|\beta^*\|^2 + C\sum_i p_i[1 - y_i\langle\beta^*, \phi_i\rangle]_+ \leq \\ &\qquad \frac{1}{2}\left\|\frac{1}{\gamma}\beta_0\right\|^2 + C\sum_i p_i[1 - y_i\left\langle\frac{1}{\gamma}\beta_0, \phi_i\right\rangle]_+ \\ &= \frac{1}{2\gamma^2} + C\mathbf{E}\left[[1 - y\left\langle\frac{1}{\gamma}\beta_0, \phi(x)\right\rangle]_+\right] = \frac{1}{2\gamma^2} + C\epsilon_0 \end{aligned} \tag{6.7}$$

Since both terms on the left hand side are non-negative, each of them is bounded by the right hand side, and in particular:

$$C\sum_i p_i[1 - y_i\langle\beta^*, \phi_i\rangle]_+ \leq \frac{1}{2\gamma^2} + C\epsilon_0 \tag{6.8}$$

Dividing by C we get a bound on the average hinge-loss of the predictor β^*, relative to a margin of one:

$$\mathbf{E}[[1 - y\langle\beta^*, \phi(x)\rangle]_+] \leq \frac{1}{2C\gamma^2} + \epsilon_0 \tag{6.9}$$

We now use the fact that β^* can be written as $\beta^* = \sum_i \alpha_i^* y_i \phi_i$ with $0 \leq \alpha_i^* \leq C p_i$. Let us consider the weights

$$w_i = w(x_i) = \alpha_i^*/(A p_i) \leq 1 \tag{6.10}$$

So, $w_i \leq \frac{C}{A}$ and $\mathbf{E}[w] = \frac{\sum_i \alpha_i^*}{A}$. Furthermore, since we have no duality gap we also have

$$\sum_i \alpha_i^* - \frac{1}{2}\|\beta^*\|^2 = \frac{1}{2}\|\beta^*\|^2 + C\sum_i p_i[1 - y_i\langle\beta^*, \phi_i\rangle]_+,$$

so $\sum_i \alpha_i^* \leq \frac{1}{\gamma^2} + C\epsilon_0$.

So, we have for every x, y:

$$\begin{aligned} y\mathbf{E}_{x',y'}[w(x')y'K(x,x')] &= y\sum_i p_i w(x_i) y_i K(x, x_i) \\ &= y\sum_i p_i \alpha_i^* y_i K(x, x_i)/(A p_i) \\ &= y\sum_i \alpha_i^* y_i \langle\phi_i, \phi(x)\rangle/A \\ &= y\langle\beta^*, \phi(x)\rangle/A \end{aligned}$$

Multiplying by A and using (6.9):

$$\begin{aligned} &\mathbf{E}_{x,y}[\,[\,1 - Ay\mathbf{E}_{x',y'}[w(x')y'K(x,x')]\,]_+\,] \\ &= \mathbf{E}_{x,y}[\,[\,1 - y\langle\beta^*, \phi(x)\rangle\,]_+\,] \leq \frac{1}{2C\gamma^2} + \epsilon_0 \end{aligned} \tag{6.11}$$

This holds for any A and C such that $\left(\frac{1}{\gamma^2} + C\epsilon_0\right)\frac{1}{A} \leq 1$, and describes the average hinge-loss relative to margin $1/A$. We also have the constraint $\frac{C}{A} \leq M$. Choosing $M = \frac{1}{2\epsilon_1+\epsilon_0}$, $A = \frac{1+\epsilon_0/2\epsilon_1}{\gamma^2}$, we set $C = 1/(2\epsilon_1\gamma^2)$ and get an average hinge-loss of $\epsilon_0 + \epsilon_1$,

$$\mathbf{E}_{x,y}\big[\,[\,1 - y\mathbf{E}_{x',y'}[w(x')y'K(x,x')]/(2\epsilon_1\gamma^2)\,]_+\,\big] \leq \epsilon_0 + \epsilon_1 \tag{6.12}$$

as desired.

This establishes that if K is (ϵ_0, γ)-good kernel in hinge loss then it is also a strongly $(\epsilon_0+\epsilon_1, \frac{\gamma^2}{1+\epsilon_0/2\epsilon_1}, \frac{1}{2\epsilon_1+\epsilon_0})$-good similarity in hinge loss, for any $\epsilon_1 > 0$, at least for finite discrete distributions.

To extend the result also to non-discrete distributions, we can consider the variational "infinite SVM" problem and apply the same arguments, as in (Srebro, 2007). ■

We can now use the hinge-loss correspondence to get a similar result for the margin-violation definitions:

Theorem 21 *If K is (ϵ_0, γ)-good kernel for a learning problem (with deterministic labels), then it is also a strongly $(\epsilon_0 + \epsilon_1, \gamma^2/2, \frac{1}{(1-\epsilon_0)\epsilon_1})$-good similarity function for the learning problem, for any $\epsilon_1 > 0$.*

Proof: If K is $(0, \gamma)$-good as a kernel, it is also $(0, \gamma)$ good as a kernel in hinge loss, and we can apply Theorem 20 to obtain that K is also $(\epsilon_0/2, \gamma_1, \tau_1)$-good, where $\gamma_1 = \gamma^2$ and $\tau_1 = 1/\epsilon_1$. We can then bound the number of margin violations at $\gamma_2 = \gamma_1/2$ by half the hinge loss at margin γ_1 to obtain the desired result.

If K is only (ϵ, γ)-good as a kernel, we follow a similar procedure to that described in (Srebro, 2007), and consider a distribution conditioned only on those places where there is no error. Returning to the original distribution, we must scale the weights up by an amount proportional to the probability of the event we conditioned on (i.e. the probability of no margin violation). This yields the desired bound. ■

7 Learning with Multiple Similarity Functions

Suppose that rather than having a single similarity function, we were instead given n functions $K_1, ..., K_n$, and our hope is that some convex combination of them will satisfy Definition 6. Is this sufficient to be able to learn well? (Note that a convex combination of similarity functions is guaranteed to have range $[-1, 1]$ and so be a legal similarity function.) The following generalization of Theorem 8 shows that this is indeed the case. (The analog of Theorem 11 can be derived similarly.)

Theorem 22 *Suppose $K_1, \ldots, K_n$ are similarity functions such that some (unknown) convex combination of them is (ϵ, γ, τ)-good. For any $\delta > 0$, let $S = \{x'_1, x'_2, \ldots, x'_d\}$ be a sample of size $d = 16\frac{\log(1/\delta)}{\tau\gamma^2}$ drawn from P. Consider the mapping $\phi^S : X \to \mathbb{R}^{nd}$ defined as follows: $\phi^S{}_i(x) = (K_1(x, x'_1), \ldots, K_n(x, x'_1), \ldots, K_1(x, x'_d), \ldots, K_n(x, x'_d))$.*

With probability at least $1 - \delta$ over the random sample S, the induced distribution $\phi^S(P)$ in R^{nd} has a separator of error at most $\epsilon + \delta$ at L_1, L_∞ margin at least $\gamma/2$.

Proof: Let $K = \alpha_1 K_1 + \ldots + \alpha_n K_n$ be an (ϵ, γ, τ)-good convex-combination of the K_i. By Theorem 8, had we instead performed the mapping: $\tilde{\phi}^S : X \to R^d$ defined as

$$\tilde{\phi}^S(x) = (K(x, x'_1), \ldots, K(x, x'_d)),$$

then with probability $1 - \delta$, the induced distribution $\tilde{\phi}^S(P)$ in R^d would have a separator of error at most $\epsilon + \delta$ at margin at least $\gamma/2$. Let $\hat{\beta}$ be the vector corresponding to such a separator in that space. Now, let us convert $\hat{\beta}$ into a vector in R^{nd} by replacing each coordinate $\hat{\beta}_j$ with the n values $(\alpha_1\hat{\beta}_j, \ldots, \alpha_n\hat{\beta}_j)$. Call the resulting vector $\tilde{\beta}$. Notice that by design, for any x we have $\left\langle \tilde{\beta}, \phi^S(x) \right\rangle = \left\langle \hat{\beta}, \tilde{\phi}^S(x) \right\rangle$. Furthermore, $\left\|\tilde{\beta}\right\|_1 = \left\|\hat{\beta}\right\|_1$. Thus, the vector $\tilde{\beta}$ under distribution $\phi^S(P)$ has the same properties as the vector $\hat{\beta}$ under $\tilde{\phi}^S(P)$. This implies the desired result. ■

Note that we get significantly better bounds here than in (Balcan & Blum, 2006), since the margin does not drop by a factor of $\frac{1}{\sqrt{n}}$.

8 Conclusions

We provide a new notion of a "good similarity function" that we prove is strictly more powerful than the traditional notion of a large-margin kernel. Our new notion relies upon L_1 regularized learning, and our separation result is related to a separation result between what is learnable with L_1 vs. L_2 regularization. In a lower bound of independent interest, we show that if C is a class of n pairwise uncorrelated functions, then *no* kernel is (ϵ, γ)-good in hinge-loss for all $f \in C$ even for $\epsilon = 0.5$ and $\gamma = 8/\sqrt{n}$.

It would be interesting to explore whether the lower bound could be extended to cover *margin violations* with a constant error rate $\epsilon > 0$ rather than only hinge-loss. In addition, it would be particularly interesting to develop even broader natural notions of good similarity functions, that allow for functions that are not positive-semidefinite and yet provide even better kernel-to-similarity translations (e.g., not squaring the margin parameter).

Acknowledgments: We would like to thank Manfred Warmuth and Hans-Ulrich Simon for helpful discussions. This work was supported in part by the National Science Foundation under grant CCF-0514922, by an IBM Graduate Fellowship, and by a Google Research Grant.

References

Arora, S., Babai, L., Stern, J., & Sweedyk, Z. (1997). The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, *54*, 317 – 331.

Balcan, M.-F., & Blum, A. (2006). On a theory of learning with similarity functions. *Proceedings of the 23rd International Conference on Machine Learning*.

Bartlett, P. L., & Mendelson, S. (2003). Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, *3*, 463–482.

Ben-David, S., Eiron, N., & Simon, H.-U. (2003). Limitations of learning via embeddings in euclidean half-spaces. *The Journal of Machine Learning Research*, *3*, 441 – 461.

Benedek, G., & Itai, A. (1988). Learnability by fixed distributions. *Proc. 1st Workshop Computat. Learning Theory* (pp. 80–90).

Bennett, K. P., & Campbell, C. (2000). Support vector machines: hype or hallelujah? *SIGKDD Explor. Newsl.*, *2*, 1–13.

Blum, A., Furst, M., Jackson, J., Kearns, M., Mansour, Y., & Rudich, S. (1994). Weakly learning DNF and characterizing statistical query learning using fourier analysis. *Proceedings of the 26th Annual ACM Symposium on Theory of Computing* (pp. 253–262).

Chapelle, O., Schlkopf, B., & Zien, A. (2006). *Semi-supervised learning*. MIT Press.

Feldman, V., Gopalan, P., Khot, S., & Ponnuswami, A. (2006). New results for learning noisy parities and halfspaces. *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (pp. 563–574).

Forster, J., & Simon, H.-U. (2006). On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theoretical Computer Science*, *350*, 40–48.

Girosi, F. (1998). An equivalence between sparse approximation and support vector machines. *Neural Comput., 10,* 1455–1480.

Guigue, V., Rakotomamonjy, A., & Canu, S. (2005). Kernel basis pursuit. *Proceedings of the 16th European Conference on Machine Learning (ECML'05).* Springer.

Gunn, S. R., & Kandola, J. S. (2002). Structural modelling with sparse kernels. *Mach. Learn., 48,* 137–163.

Guruswami, V., & Raghavendra, P. (2006). Hardness of learning halfspaces with noise. *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (pp. 543–552).

Herbrich, R. (2002). *Learning kernel classifiers*. MIT Press, Cambridge.

Linial, N., Mansour, Y., & Nisan, N. (1989). Constant depth circuits, fourier transform, and learnability. *Proceedings of the Thirtieth Annual Symposium on Foundations of Computer Science* (pp. 574–579). Research Triangle Park, North Carolina.

Littlestone, N. (1989). From online to batch learning. *Proc. 2nd Annual ACM Conference on Computational Learning Theory* (pp. 269–284).

Mansour, Y. (1994). Learning boolean functions via the fourier transform. In *Theoretical advances in neural computation and learning*, 391–424.

McAllester, D. (2003). Simplified pac-bayesian margin bounds. *Proceedings of the 16th Conference on Computational Learning Theory*.

Mitchell, T. (2006). The discipline of machine learning. *CMU-ML-06 108*.

Osuna, E. E., & Girosi, F. (1999). Reducing the run-time complexity in support vector machines. In *Advances in kernel methods: support vector learning*, 271–283. Cambridge, MA, USA: MIT Press.

Roth, V. (2001). Sparse kernel regressors. *ICANN '01: Proceedings of the International Conference on Artificial Neural Networks* (pp. 339–346). London, UK: Springer-Verlag.

Scholkopf, B., & Smola, A. J. (2002). *Learning with kernels. support vector machines, regularization, optimization, and beyond*. MIT University Press, Cambridge.

Scholkopf, B., Tsuda, K., & Vert, J.-P. (2004). *Kernel methods in computational biology*. MIT Press.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Singer, Y. (2000). Leveraged vector machines. *Advances in Neural International Proceedings System 12*.

Smola, A. J., & Schölkopf, B. (2002). *Learning with kernels*. MIT Press.

Srebro, N. (2007). How Good is a Kernel as a Similarity Function. *COLT*.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res., 1,* 211–244.

Vincent, P., & Bengio, Y. (2002). Kernel matching pursuit. *Mach. Learn., 48,* 165–187.

Warmuth, M. K., & Vishwanathan, S. V. N. (2005). Leaving the span. *Proceedings of the Annual Conference on Learning Theory*.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *J. Mach. Learn. Res., 2,* 527–550.

A Kernels and Similarity Functions

Theorem 23 *If K is an (ϵ, γ)-good similarity function under Definitions 4 and 5, then K is also an $(\epsilon, \gamma, \gamma)$-good similarity function under Definitions 6 and 7, respectively.*

Proof: If we set $\Pr(R(x) \mid x) = w(x)$, we get that in order for any point x to fulfill equation (2.1), we must have

$$\Pr(R(x)) = \mathbf{E}[w(x)] \geq \mathbf{E}[\ell\ell' w(x')K(x, x')] \geq \gamma.$$

Furthermore, for any x, ℓ for which (2.1) is satisfied, we have

$$\begin{aligned}\mathbf{E}[\ell\ell' K(x, x')|R(x')] &= \mathbf{E}[\ell\ell' K(x, x')w(x')]/\Pr(R(x)) \\ &\geq \mathbf{E}[\ell\ell' K(x, x')w(x')] \geq \gamma \quad \text{(A.1)}\end{aligned}$$

. ■

Theorem 24 *If K is an (ϵ, γ, τ)-good similarity function under Definitions 6 and 7, then K is an $(\epsilon, \gamma\tau)$-good similarity function under Definitions 4 and 5 (respectively).*

Proof: Setting $w(x) = \Pr(R(x) \mid x)$ we have for any x, ℓ satisfying (3.1) that

$$\begin{aligned}\mathbf{E}[\ell\ell' K(x, x')w(x')] &= \mathbf{E}[\ell\ell' K(x, x')R(x')] = \\ &\mathbf{E}[\ell\ell' K(x, x')|R(x')]\Pr(R(x')) \geq \gamma\tau. \quad \text{(A.2)}\end{aligned}$$

A similar calculation establishes the correspondence for the hinge loss. ■

We show in the following that a kernel good as a similarity function is also good as a kernel.

Theorem 25 *If K is a valid kernel function, and is (ϵ, γ, τ)-good similarity for some learning problem, then it is also (ϵ, γ)-kernel-good for the learning problem. If K is (ϵ, γ, τ)-good similarity in hinge loss, then it is also (ϵ, γ)-kernel-good in hinge loss.*

Proof: Consider a similarity function K that is a valid kernel, i.e. $K(x, x') = \langle \phi(x), \phi(x') \rangle$ for some mapping ϕ of x to a Hilbert space $\mathcal{H}$. For any input distribution and any probabilistic set of reasonable points R of the input we will construct a linear predictor $\beta_w \in \mathcal{H}$, with $\|\beta_w\| \leq 1$, such that similarity-based predictions using R are the same as the linear predictions made with β_R.

Define the following linear predictor $\beta_R \in \mathcal{H}$:

$$\beta_R = \mathbf{E}_{x'}[\ell(x')\phi(x')|R(x')]. \tag{A.3}$$

The predictor β_w has norm at most:

$$\begin{aligned}\|\beta_R\| = \|\mathbf{E}_{x'}[\ell(x')\phi(x')|R(x')]\| &\leq \max_{x'} \|\ell(x')\phi(x')\| \\ &\leq \max \|\phi(x')\| = \max \sqrt{K(x',x')} \leq 1 \end{aligned} \tag{A.4}$$

where the second inequality follows from $|\ell(x')| \leq 1$.

The predictions made by β_R are:

$$\begin{aligned}\langle \beta_R, \phi(x)\rangle = \langle \mathbf{E}_{x'}[\ell(x')\phi(x')|R(x')], \phi(x)\rangle = \\ \mathbf{E}_{x'}[\ell(x')\langle\phi(x'), \phi(x)\rangle|R(x')] = \mathbf{E}_{x'}[\ell(x'))K(x,x')|R(x')]\end{aligned} \tag{A.5}$$

That is, using β_R is the same as using similarity-based prediction with R. In particular, the margin violation rate, as well as the hinge loss, with respect to any margin γ, is the same for predictions made using either R or β_R. This is enough to establish Theorem 25: If K is (ϵ, γ)-good (perhaps for to the hinge-loss), there exists some valid R that yields margin violation error rate (resp. hinge loss) at most ϵ with respect to margin γ, and so β_R yields the same margin violation (resp. hinge loss) with respect to the same margin, establishing K is (ϵ, γ)-kernel-good (resp. for the hinge loss). ■