How Good is a Kernel When Used as a Similarity Measure?

Nathan Srebro

Toyota Technological Institute-Chicago IL, USA
IBM Haifa Research Lab, ISRAEL
nati@uchicago.edu

Abstract. Recently, Balcan and Blum [1] suggested a theory of learning based on general similarity functions, instead of positive semi-definite kernels. We study 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, which was left open by Balcan and Blum. We provide a significantly improved bound on how good a kernel function is when used as a similarity function, and extend the result also to the more practically relevant hinge-loss rather then zero-one-error-rate. Furthermore, we show that this bound is tight, and hence establish that there is in-fact a real gap between the traditional kernel-based notion of margin and the newer similarity-based notion.

1 Introduction

A common contemporary approach in machine learning is to encode prior knowledge about objects using a *kernel*, specifying the inner products between implicit high-dimensional representations of objects. Such inner products can be viewed as measuring the *similarity* between objects. In-fact, many generic kernels (e.g. Gaussian kernels), as well as very specific kernels (e.g. Fisher kernels [2] and kernels for specific structures such as [3]), describe different notions of similarity between objects, which do not correspond to any intuitive or easily interpretable high-dimensional representation. However, not every mapping of pairs of objects to “similarity values” is a valid kernel.

Recently, Balcan and Blum [1] proposed an alternative theory of learning, which is based on a more general notion of similarity functions between objects, which unlike valid kernel functions, need not be positive semi-definite. Balcan and Blum provide a definition for a separation *margin* of a classification problem under a general similarity measure and present learning methods with guarantees that parallel the familiar margin-based guarantees for kernel methods.

It is interesting to study what this alternative theory yields for similarity functions which are in-fact valid kernel functions. Does the similarity-based theory subsume the kernel-based theory without much deterioration of guarantees? Or can the kernel-based theory provide better results for functions which are in-fact positive semi-definite. To answer these questions, one must understand how a kernel-based margin translates to a similarity-based margin. Balcan and

Blum showed that if an input distribution can be separated, in the kernel sense, with margin γ and error rate ϵ_0 (i.e. ϵ_0 of the inputs are allowed to violate the margin), then viewing the kernel mapping as a similarity measure, for any $\epsilon_1 > 0$, the target distribution can be separated with similarity-based margin[1] $\frac{\gamma\epsilon_1}{96/\gamma^2 - 32\log\epsilon_1} = \tilde{\Theta}\left(\epsilon_1\gamma^3\right)$ and error rate $8\epsilon_0/\gamma + \epsilon_1$. Although this does establish that good kernels can also be used as similarity measures, in the Blum and Balcan sense, there is a significant deterioration in the margin yielding a significant deterioration in the learning guarantee. The tightness of this relationship, or a possible improved bound, was left unresolved. Also, this result of Balcan and Blum refers only to a zero-one error-rate, which does not yield efficient learning algorithms. Guarantees referring to the hinge-loss are desirable.

Here, we resolve this question by providing an improved bound, with a simpler proof, and establishing its tightness. We show that:

- If an input distribution can be separated, in the kernel sense, with margin γ and error rate ϵ_0, then for any $\epsilon_1 > 0$, it can also be separated by the kernel mapping viewed as a similarity measure, with similarity-based margin $\frac{1}{2}(1-\epsilon_0)\epsilon_1\gamma^2$ and error rate $\epsilon_0 + \epsilon_1$.
- We also obtain a similar bound in terms of the average hinge loss, instead of the margin violation error rate: If for a target distribution we can achieve, in the kernel sense, average hinge loss of ϵ_0 for margin γ, then for any $\epsilon_1 > 0$, we can also achieve average hinge loss of $\epsilon_0 + \epsilon_1$ for margin $2\epsilon_1\gamma^2$, when the kernel mapping is used as a similarity measure. A result in terms of the hinge-loss is perhaps more practical, since for computational reasons, we usually minimize the hinge-loss rather then error rate.
- The above bounds are tight, up to a factor of sixteen: We show, for any $\gamma < \frac{1}{2}$ and ϵ_1, a specific kernel function and input distribution that can be separated with margin γ and no errors in the kernel sense, but which can only be separated with margin at most $32\epsilon_1\gamma^2$ in the similarity sense, if we require hinge loss less than ϵ_1 or error-rate less than $4\epsilon_1$, when using the same kernel mapping as a similarity measure.

In the next Section we formally present the framework in which we work and remind the reader of the definitions and results of Balcan and Blum. We then state our results (Section 3) and prove them (Sections 4 and 5).

2 Setup

We briefly review the setting used by Balcan and Blum [1], which we also use here.

We consider input distributions (X, Y) over $\mathcal{X} \times \{\pm 1\}$, where $\mathcal{X}$ is some abstract object space. As in Balcan and Blum [1], we consider only *consistent* input distributions in which the label Y is a *deterministic* function of X. We can think of such input distributions as a distributions over $\mathcal{X}$ and a deterministic mapping $y(x)$.

[1] The $\tilde{\Theta}(\cdot)$ and $\tilde{\mathcal{O}}(\cdot)$ notations hide logarithmic factors.

A *kernel function* is a mapping $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which there exists an (implicit) feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ of objects into an (implicit) Hilbert space $\mathcal{H}$ such that $K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$. See, e.g., Smola and Schölkopf [4] 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 $x \in \mathcal{X}$. Kernalized large-margin classification relies on the existence of a large margin linear separator for the input distribution, in the Hilbert space implied by K. This is captured by the following definition of when a kernel function is *good* for an input distribution:

Definition 1. *A kernel K is* (ϵ, γ)**-kernel-good** *for an input distribution if there exists a classifier $\beta \in \mathcal{H}$, $\|\beta\| = 1$, such that* $\Pr(\, Y\langle \beta, \phi(X) \rangle < \gamma \,) \leq \epsilon$. *We say β has margin-γ-error-rate* $\Pr(\, Y\langle \beta, \phi(X) \rangle < \gamma \,)$.

Given a kernel that is (ϵ, γ)-kernel-good (for some unknown source distribution), a predictor with error rate at most $\epsilon + \epsilon_{\text{acc}}$ (on the source distribution) can be learned (with high probability) from a sample of $\tilde{\mathcal{O}}\big((\epsilon + \epsilon_{\text{acc}})/(\gamma^2 \epsilon_{\text{acc}}^2)\big)$ examples (drawn independently from the source distribution) by minimizing the number of margin γ violations on the sample [5]. However, minimizing the number of margin violations on the sample is a difficult optimization problem. Instead, it is common to minimize the so-called *hinge loss* relative to a margin:

Definition 2. *A kernel K is* (ϵ, γ)**-kernel-good in hinge-loss** *for an input distribution if there exists a classifier $\beta \in \mathcal{H}$, $\|\beta\| = 1$, such that*

$$\mathbf{E}[[1 - Y\langle \beta, \phi(X) \rangle / \gamma]_+] \leq \epsilon,$$

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 (with high probability) from a sample of $\mathcal{O}\big(1/(\gamma^2 \epsilon_{\text{acc}}^2)\big)$ examples by minimizing the average hinge loss relative to margin γ on the sample [6].

A *similarity function* is any symmetric mapping $K : \mathcal{X} \times \mathcal{X} \to [-1, +1]$. In particular, a (properly normalized) kernel function is also a similarity function. Instead of functionals in an implicit Hilbert space, similarity-based predictors are given in terms of a *weight function* $w : \mathcal{X} \to [0, 1]$. The classification margin of (x, y) is then defined as [1]:

$$\begin{aligned} \mathbf{E}_{X',Y'}[w(X')Y'K(x, X')|y = Y'] - \mathbf{E}_{X',Y'}[w(X')Y'K(x, X')|y \neq Y'] \\ = y\mathbf{E}_{X',Y'}[w(X')Y'K(x, X')/p(Y')] \end{aligned} \tag{1}$$

where $p(Y')$ is the marginal probability of the label, i.e. the prior. We choose here to stick with this definition used by Balcan and Blum. All our results apply (up to a factor for 1/2) also to a weaker definition, dropping the factor $1/p(Y')$ from definition of the classification margin (1).

We are now ready to define when a similarity function is good for an input distribution:

Definition 3. *A similarity function K is* (ϵ, γ)**-similarity-good** *for an input distribution if there exists a mapping $w : \mathcal{X} \to [0, 1]$ such that:*

$$\Pr_{X,Y}\left(Y\mathbf{E}_{X',Y'}[w(X')Y'K(X, X')/p(Y')] < \gamma \right) \leq \epsilon.$$

Balcan and Blum showed how, given a similarity function that is (ϵ, γ)-similarity-good, a predictor with error at most $\epsilon + \epsilon_{\text{acc}}$ can be learned (with high probability) from a sample of $\tilde{\mathcal{O}}\big((\epsilon + \epsilon_{\text{acc}})/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ examples. This is done by first using $\tilde{\mathcal{O}}\big(1/\gamma^2\big)$ positive and $\tilde{\mathcal{O}}\big(1/\gamma^2\big)$ negative examples to construct an explicit feature map ϕ which is $(\epsilon + \epsilon_{\text{acc}}/2, \gamma/4)$-kernel-good (that is, the inner product in this space is a good kernel) [1, Theorem 2], and then searching for a margin $\gamma/4$ linear separator in this space minimizing the number of margin violations. As mentioned before, this last step (minimizing margin violations) is a difficult optimization problem. We can instead consider the hinge-loss:

Definition 4. *A similarity function K is* (ϵ, γ)**-similarity-good in hinge loss** *for an input distribution if there exists a mapping $w : \mathcal{X} \to [0, 1]$ such that:*

$$\mathbf{E}_{X,Y}[[1 - Y\mathbf{E}_{X',Y'}[w(X')Y'K(X, X')/p(Y')]/\gamma]_+] \leq \epsilon.$$

Using the same approach as above, given a similarity function that is (ϵ, γ)-similarity-good in hinge loss, a predictor with error at most $\epsilon + \epsilon_{\text{acc}}$ can be efficiently learned (with high probability) from a sample of $\mathcal{O}\big(1/(\gamma^2\epsilon_{\text{acc}}^2)\big)$ examples, where this time in the second stage the hinge loss, rather then the number of margin violations, is minimized.

We see, then, that very similar learning guarantees can be obtained by using mappings that are (ϵ, γ)-kernel-good or (ϵ, γ)-similarity-good. A natural question is then, whether a kernel that is (ϵ, γ)-kernel-good is also $(\epsilon, \Omega\gamma)$-similarity-good. A positive answer would indicate that learning guarantees based on similarity-goodness subsume the more restricted results based on kernel-goodness (up to constant factors). However, a negative result would indicate that for a mapping that is a valid kernel (i.e. is positive semi-definite), the theory of kernel-based learning provides stronger guarantees than those that can be established using Balcan and Blum's learning methods and guarantees based on similarity goodness (it is still possible that stronger similarity-based guarantees might be possible using a different learning approach).

3 Summary of Results

Considering the question of whether the theory of learning with similarity function subsumes the theory of learning with kernels, Balcan and Blum showed [1, Theorem 4] that a kernel that is (ϵ_0, γ)-kernel-good for a (consistent) input distribution, is also $(8\epsilon_0/\gamma + \epsilon_1, \frac{\gamma\epsilon_1}{96/\gamma^2 - 32\log\epsilon_1})$-similarity-good for the input distribution, for any $\epsilon_1 > 0$. This result applies only to margin violation goodness, and not to the more practically useful hinge-loss notion of goodness. The result still leaves a large gap even for the margin violation case, as the margin is decreased from γ to $\tilde{\Theta}\big(\epsilon_1\gamma^3\big)$, and the error is increased by both an additive factor of ϵ_1 and a multiplicative factor of $8/\gamma$.

First, we improve on this result, obtaining a better guarantee on similarity-goodness based on kernel-goodness, that applies both for margin-violations and for hinge-loss:

Theorem 1 (Main Result, Margin Violations). *If K is (ϵ_0, γ)-kernel-good for some (consistent) input distribution, then it is also $(\epsilon_0 + \epsilon_1, \frac{1}{2}(1-\epsilon_0)\epsilon_1\gamma^2)$-similarity-good for the distribution, for any $\epsilon_1 > 0$.*

Note that in any useful situation $\epsilon_0 < \frac{1}{2}$, and so the guaranteed margin is at least $\frac{1}{4}\epsilon_1\gamma^2$.

Theorem 2 (Main Result, Hinge Loss). *If K is (ϵ_0, γ)-kernel-good in hinge loss for some (consistent) input distribution, then it is also $(\epsilon_0 + \epsilon_1, 2\epsilon_1\gamma^2)$-similarity-good in hinge loss for the distribution, for any $\epsilon_1 > 0$.*

These guarantees still yield a significant deterioration of the margin, when considering similarity-goodness as opposed to kernel-goodness. However, we establish that this is the best that can be hoped for by presenting examples of kernels for which these guarantees are tight (up to a small multiplicative factor):

Theorem 3 (Tightness, Margin Violations). *For any $0 < \gamma < \sqrt{1/2}$ and any $0 < \epsilon_1 < 1/2$, there exists an input distribution and a kernel function K, which is $(0, \gamma)$-kernel-good for the input distribution, but which is only $(\epsilon_1, 8\epsilon_1\gamma^2)$-similarity-good. That is, it is not (ϵ_1, γ')-similarity-good for any $\gamma' > 8\epsilon_1\gamma^2$.*

Theorem 4 (Tightness, Hinge Loss). *For any $0 < \gamma < \sqrt{1/2}$ and any $0 < \epsilon_1 < 1/2$, there exists an input distribution and a kernel function K, which is $(0, \gamma)$-kernel-good in hinge loss for the input distribution, but which is only $(\epsilon_1, 32\epsilon_1\gamma^2)$-similarity-good in hinge loss.*

4 An Improved Guarantee

We are now ready to prove Theorems 1 and 2. We will consider a kernel function that is (ϵ_0, γ)-kernel-good and show that it is also good as a similarity function. We begin, in Section 4.1, with goodness in hinge-loss, and prove Theorem 2, which can be viewed as a more general result. Then, in Section 4.2, we prove Theorem 1 in terms of the margin violation error rate, by using the hinge-loss as a bound on the error rate.

In either case, our proof is based on the representation of the optimal SVM solution in terms of the dual optimal solution.

4.1 Proof of Theorem 2: Goodness in hinge-loss

We consider consistent input distributions, in which Y is a deterministic function of X. For simplicity of presentation, we first consider finite discrete distributions, where:

$$\Pr(\,(X,Y) = (x_i, y_i)\,) = p_i \tag{2}$$

for $i = 1 \dots n$, 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 for our input distribution. 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:

$$\text{minimize } \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^{n} p_i [1 - y_i \langle \beta, \phi_i \rangle]_+ \tag{3}$$

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

$$\begin{aligned} \text{maximize } & \sum_i \alpha_i - \frac{1}{2} \sum_{ij} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \\ \text{subject to } & \; 0 \leq \alpha_i \leq C p_i \end{aligned} \tag{4}$$

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

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 (3), being the optimum solution, then satisfies:

$$\begin{aligned} \frac{1}{2} \|\beta^*\|^2 + C \sum_i p_i [1 - y_i \langle \beta^*, \phi_i \rangle]_+ \leq & \\ & \frac{1}{2} \left\| \frac{1}{\gamma} \beta_0 \right\|^2 + C \sum_i p_i [1 - y_i \left\langle \frac{1}{\gamma} \beta_0, \phi_i \right\rangle]_+ \\ & = \frac{1}{2\gamma^2} + C \mathbf{E} \left[[1 - Y \left\langle \frac{1}{\gamma} \beta_0, \phi(X) \right\rangle]_+ \right] = \frac{1}{2\gamma^2} + C \epsilon_0 \end{aligned} \tag{5}$$

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

$$C \sum_i p_i [1 - y_i \langle \beta^*, \phi_i \rangle]_+ \leq \frac{1}{2\gamma^2} + C \epsilon_0 \tag{6}$$

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

$$\mathbf{E}[[1 - Y \langle \beta^*, \phi(X) \rangle]_+] \leq \frac{1}{2C\gamma^2} + \epsilon_0 \tag{7}$$

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

$$w_i = w(x_i) = \alpha_i^* p(y_i) / (C p_i) \leq p(y_i) \leq 1 \tag{8}$$

we have for every x, y:

$$\begin{aligned} y\mathbf{E}_{X',Y'}[w(X')Y'K(x,X')/p(Y')] &= y\sum_i p_i w(x_i) y_i K(x,x_i)/p(y_i) \\ &= y\sum_i p_i \alpha_i^* p(y_i) y_i K(x,x_i)/(Cp_i p(y_i)) \\ &= y\sum_i \alpha_i^* y_i \langle \phi_i, \phi(x)\rangle / C = y\langle \beta^*, \phi(x)\rangle / C \end{aligned} \tag{9}$$

Multiplying by C and using (7):

$$\begin{aligned} &\mathbf{E}_{X,Y}[\,[\,1 - CY\mathbf{E}_{X',Y'}[w(X')Y'K(X,X')/p(Y')]\,]_+\,] \\ &\qquad = \mathbf{E}_{X,Y}[\,[\,1 - Y\langle \beta^*, \phi(X)\rangle\,]_+\,] \le \frac{1}{2C\gamma^2} + \epsilon_0 \end{aligned} \tag{10}$$

This holds for any C, and describes the average hinge-loss relative to margin $1/C$. To get an average hinge-loss of $\epsilon_0 + \epsilon_1$, we set $C = 1/(2\epsilon_1\gamma^2)$ and get:

$$\mathbf{E}_{X,Y}\left[\,[\,1 - Y\mathbf{E}_{X',Y'}[w(X')Y'K(X,X')/p(Y')]/(2\epsilon_1\gamma^2)\,]_+\,\right] \le \epsilon_0 + \epsilon_1 \tag{11}$$

This establishes that K is $(\epsilon_0 + \epsilon_1, 2\epsilon_1\gamma^2)$-similarity-good in hinge-loss.

Non-discrete input distribution The same arguments apply also in the general (not necessarily discrete) case, except that this time, instead of a fairly standard (weighted) SVM problem, we must deal with a variational optimization problem, where the optimization variable is a random variable (a function from the sample space to the reals). We will present the dualization in detail.

We consider the primal objective

$$\text{minimize } \frac{1}{2}\|\beta\|^2 + C\mathbf{E}_{Y,\phi}[[1 - Y\langle \beta, \phi\rangle]_+] \tag{12}$$

where the expectation is w.r.t. the input distribution, with $\phi = \phi(X)$ here and throughout the rest of this section. We will rewrite this objective using explicit slack, in the form of a random variable ξ, which will be a variational optimization variable:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\|\beta\|^2 + C\mathbf{E}[\xi] \\ &\text{subject to } \Pr(\,1 - y\langle \beta, \phi\rangle - \xi \le 0\,) = 1 \\ &\qquad\qquad \Pr(\,\xi \ge 0\,) = 1 \end{aligned} \tag{13}$$

In the rest of this section all our constraints will implicitly be required to hold with probability one. We will now introduce the dual variational optimization variable α, also a random variable over the same sample space, and write the problem as a saddle problem:

$$\begin{aligned} &\min_{\beta,\xi} \max_{\alpha} \frac{1}{2}\|\beta\|^2 + C\mathbf{E}[\xi] + \mathbf{E}[\alpha(1 - Y\langle \beta, \phi\rangle - \xi)] \\ &\text{subject to } \quad \xi \ge 0 \quad \alpha \ge 0 \end{aligned} \tag{14}$$

Note that this choice of Lagrangian is a bit different than the more standard Lagrangian leading to (4). Convexity and the existence of a feasible point in the dual interior allows us to change the order of maximization and minimization without changing the value of the problem [7]. Rearranging terms we obtaining the equivalent problem:

$$\begin{gathered}\max_\alpha \min_{\beta,\xi} \frac{1}{2}\left\|\beta\right\|^2 - \langle \mathbf{E}[\alpha Y \phi], \beta\rangle + \mathbf{E}[\xi(C-\alpha)] + \mathbf{E}[\alpha] \\ \text{subject to} \quad \xi \geq 0, \quad \alpha \geq 0\end{gathered} \tag{15}$$

Similarly to the finite case, we see that the minimum of the minimization problem is obtained when $\beta = \mathbf{E}[\alpha Y \phi]$ and that it is finite when $\alpha \leq C$ almost surely, yielding the dual:

$$\begin{gathered}\text{maximize } \mathbf{E}[\alpha] - \frac{1}{2}\mathbf{E}[\alpha Y \alpha' Y K(X, X')] \\ \text{subject to} \quad 0 \leq \alpha \leq C\end{gathered} \tag{16}$$

where (X, Y, α) and (X', Y', α') are two independent draws from the same distribution. The primal optimum can be expressed as $\beta^* = \mathbf{E}[\alpha^* Y \phi]$, where α^* is the dual optimum. We can now apply the same arguments as in (5),(6) to get (7). Using the weight mapping

$$w(x) = \mathbf{E}[\alpha^* | x]\, p(y(x)) \,/\, C \leq 1 \tag{17}$$

we have for every x, y:

$$\begin{aligned} y\mathbf{E}_{X',Y'}[w(X')Y'K(x,X')/p(Y')] &= y\langle \mathbf{E}_{X',Y',\alpha'}[\alpha' Y' X'], x\rangle / C \\ &= y\langle \beta^*, \phi(x)\rangle / C. \end{aligned} \tag{18}$$

From here we can already get (10) and setting $C = 1/(2\epsilon_1\gamma^2)$ we get (11), which establishes Theorem 2 for any input distribution.

4.2 Proof of Theorem 1: Margin-violation goodness

We will now turn to guarantees on similarity-goodness with respect to the margin violation error-rate. We base these on the results for goodness in hinge loss, using the hinge loss as a bound on the margin violation error-rate. In particular, a violation of margin $\gamma/2$ implies a hinge-loss at margin γ of at least $\frac{1}{2}$. Therefore, twice the average hinge-loss at margin γ is an upper bound on the margin violation error rate at margin $\gamma/2$.

The kernel-separable case, i.e. $\epsilon_0 = 0$, is simpler, and we consider it first. Having no margin violations implies zero hinge loss. And so if a kernel K is $(0, \gamma)$-kernel-good, it is also $(0, \gamma)$-kernel-good in hinge loss, and by Theorem 2 it is $(\epsilon_1/2, 2(\epsilon_1/2)\gamma^2)$-similarity-good in hinge loss. Now, for any $\epsilon_1 > 0$, by bounding the margin $\frac{1}{2}\epsilon_1\gamma^2$ error-rate by the $\epsilon_1\gamma^2$ average hinge loss, K is $(\epsilon_1, \frac{1}{2}\epsilon_1\gamma^2)$-similarity-good, establishing Theorem 1 for the case $\epsilon_0 = 0$.

We now return to the non-separable case, and consider a kernel K that is (ϵ_0, γ)-kernel-good, with some non-zero error-rate ϵ_0. Since we cannot bound the hinge loss in terms of the margin-violations, we will instead consider a modified input distribution where the margin-violations are removed.

Since we will be modifying the input distribution, and so potentially also the label marginals, it will be simpler for us to use a definition of similarity-based margin that avoids the factor $1/p(Y')$. Therefore, in this Section, we will refer to similarity-goodness where the classification margin of (x, y) is given by:

$$y\mathbf{E}_{X',Y'}[w(X')Y'K(x, X')]. \tag{19}$$

It is easy to verify, by dropping the factor $p(y_i)$ in (8) or (17), that Theorem 2, and hence also Theorem 1 for the case $\epsilon_0 = 0$, hold also under this definition. Furthermore, if a kernel is (ϵ, γ)-good under this definition, then multiplying the label marginals into the weights establishes that it is also (ϵ, γ)-good under the definitions in Section 2.

Let β^* be the linear classifier achieving ϵ_0 margin violation error-rate with respect to margin γ, i.e. such that $\Pr(\, Y\langle\beta^*, X\rangle \geq \gamma\,) > 1 - \epsilon_0$. We will consider an input distribution which is conditioned on $Y\langle\beta^*, X\rangle \geq \gamma$. We denote this event as $\textsc{ok}(X)$ (recall that Y is a deterministic function of X). The kernel K is obviously $(0, \gamma)$-kernel-good, and so by the arguments above also $(\epsilon_1, \frac{1}{2}\epsilon_1\gamma^2)$-similarity-good, on the conditional distribution. Let w be the weight mapping achieving

$$\Pr_{X,Y}(\, Y\mathbf{E}_{X',Y'}[w(X')Y'K(X, X')|\textsc{ok}(X')] < \gamma_1 | \textsc{ok}(X)\,) \leq \epsilon_1, \tag{20}$$

where $\gamma_1 = \frac{1}{2}\epsilon_1\gamma^2$, and set $w(x) = 0$ when $\textsc{ok}(X)$ does not hold. We have:

$$\begin{aligned}
&\Pr_{X,Y}(\, Y\mathbf{E}_{X',Y'}[w(X')Y'K(X, X')] < (1 - \epsilon_0)\gamma_1\,) \\
&\leq \Pr(\text{ not } \textsc{ok}(X)\,) \\
&\qquad + \Pr(\, \textsc{ok}(X)\,) \Pr_{X,Y}(\; Y\mathbf{E}_{X',Y'}[w(X')Y'K(X, X')] < (1 - \epsilon_0)\gamma_1 \mid \textsc{ok}(X)\;) \\
&= \epsilon_0 \\
&\quad + (1{-}\epsilon_0) \Pr_{X,Y}(\, Y(1{-}\epsilon_0)\mathbf{E}_{X',Y'}[w(X')Y'K(X, X')|\textsc{ok}(X)] < (1{-}\epsilon_0)\gamma_1|\textsc{ok}(X)\,) \\
&= \epsilon_0 + (1 - \epsilon_0) \Pr_{X,Y}(\, Y\mathbf{E}_{X',Y'}[w(X')Y'K(X, X')|\textsc{ok}(X)] < \gamma_1|\textsc{ok}(X)\,) \\
&\leq \epsilon_0 + (1 - \epsilon_0)\epsilon_1 \leq \epsilon_0 + \epsilon_1
\end{aligned} \tag{21}$$

establishing that K is $(\epsilon_0{+}\epsilon_1, \gamma_1)$-similarity-good for the original (unconditioned) distribution, and yielding Theorem 1

5 Tightness

Consider a distribution on four labeled points in $\mathbb{R}^3$, which we denote x_1, x_2, x_3, x_4:

$$\begin{gathered} p(X = x_1 = (\gamma, \gamma, \sqrt{1-2\gamma^2}), Y = 1) = \frac{1}{2} - \epsilon \\ p(X = x_2 = (\gamma, -\gamma, \sqrt{1-2\gamma^2}), Y = 1) = \epsilon \\ p(X = x_3 = (-\gamma, \gamma, \sqrt{1-2\gamma^2}), Y = -1) = \epsilon \\ p(X = x_4 = (-\gamma, -\gamma, \sqrt{1-2\gamma^2}), Y = -1) = \frac{1}{2} - \epsilon \end{gathered}$$

for some (small) $0 < \gamma < \sqrt{\frac{1}{2}}$ and (small) probability $0 < \epsilon < \frac{1}{2}$. The four points are all on the unit sphere, and are clearly separated by $\beta = (1, 0, 0)$ with a margin of γ. The standard inner-product kernel is therefore $(0, \gamma)$-kernel-good on this distribution.

5.1 Margin-violation error-rate

We will show that when this kernel (the standard inner product kernel in $\mathbb{R}^3$) is used as a similarity function, the best margin that can be obtained on all four points, i.e. on at least $1 - \epsilon$ probability mass of examples, is $8\epsilon\gamma^2$.

Consider the classification margin on point x_2 with weights w (denote $w_i = w(x_i)$, and note that $p(y_i) = \frac{1}{2}$ for all i):

$$\begin{aligned} &\mathbf{E}[w(X)YK(x_2, X)/p(Y)] \\ &= 2(\frac{1}{2} - \epsilon)w_1(\gamma^2 - \gamma^2 + (1 - 2\gamma^2)) + 2\epsilon w_2(2\gamma^2 + (1 - 2\gamma^2)) \\ &- 2\epsilon w_3(-2\gamma^2 + (1 - 2\gamma^2)) - 2(\frac{1}{2} - \epsilon)w_4(-\gamma^2 + \gamma^2 + (1 - 2\gamma^2)) \\ &= 2\left((\frac{1}{2} - \epsilon)(w_1 - w_4) + \epsilon(w_2 - w_3)\right)(1 - 2\gamma^2) \; + \; 4\epsilon(w_2 + w_3)\gamma^2 \end{aligned} \tag{22}$$

If the first term is positive, we can consider the symmetric calculation

$$\begin{aligned} &- \mathbf{E}[w(X)YK(x_3, X)/p(Y)] \\ &\quad = -2\left((\frac{1}{2} - \epsilon)(w_1 - w_4) + \epsilon(w_2 - w_3)\right)(1 - 2\gamma^2) \; + \; 4\epsilon(w_2 + w_3)\gamma^2 \end{aligned} \tag{23}$$

in which the first term is negated. One of the above margins must therefore be at most

$$4\epsilon(w_2 + w_3)\gamma^2 \leq 8\epsilon\gamma^2 \tag{24}$$

This establishes Theorem 3.

5.2 Hinge loss

In the above example, suppose we would like to get an average hinge-loss relative to margin γ_1 of at most ϵ_1:

$$\mathbf{E}_{X,Y}[\,[\,1 - Y\mathbf{E}_{X',Y'}[w(X')Y'K(X,X')/p(Y')]/\gamma_1\,]_+\,] \leq \epsilon_1 \tag{25}$$

Following the arguments above, equation (24) can be used to bound the hinge-loss on at least one of the points x_2 or x_3, which, multiplied by the probability ϵ of the point, is a bound on the average hinge loss:

$$\mathbf{E}_{X,Y}[\,[\,1 - Y\mathbf{E}_{X',Y'}[w(X')Y'K(X,X')/p(Y')]/\gamma_1\,]_+\,] \geq \epsilon(1 - 8\epsilon\gamma^2/\gamma_1) \tag{26}$$

and so to get an an average hinge-loss of at most ϵ_1 we must have:

$$\gamma_1 \leq \frac{8\epsilon\gamma^2}{1 - \epsilon_1/\epsilon} \tag{27}$$

For any target hinge-loss ϵ_1, consider a distribution with $\epsilon = 2\epsilon_1$, in which case we get that the maximum margin attaining average hinge-loss ϵ_1 is $\gamma_1 = 32\epsilon_1\gamma^2$, even though we can get a hinge loss of zero at margin γ using a kernel. This establishes Theorem 4.

6 Discussion

In this paper, we studied how tightly the similarity-based theory of learning, proposed by Balcan and Blum, captures the well-studied theory of kernel-based learning. In other words, how well does a kernel-based learning guarantee translate to a similarity-based learning guarantee. We significantly improved on the bounds presented by Balcan and Blum, providing stronger, simpler, bounds that apply also in the more practically relevant case of hinge-loss minimization. However, these bounds still leave a gap between the kernel-based learning guarantees and the learning guarantee obtained when using the kernel as a similarity measure. We show that the bounds are tight, and so there is a real gap between the similarity-based theory and the kernel-based theory.

We hope that the results presented here can help us better understand similarity-based learning, and possibly suggest revisions to the theory presented by Balcan and Blum.

The quadratic increase in the margin can perhaps be avoided by using the distances, or perhaps the square root of the kernel, rather then the inner products, as a similarity function. Consider the simplest case of two points, with opposite labels and probability half, at $(\gamma, \sqrt{1-\gamma^2})$ and $(-\gamma, \sqrt{1-\gamma^2})$. The geometric margin is γ. The inner product (kernel) is only $(0, \gamma^2)$-similarity-good, but the distance function, or just the square root of the inner product, is $(0, \gamma)$-similarity-good. It would be interesting to understand what guarantees can be provided on these measures as similarity functions.

However, even if distance functions are used, the dependence on ϵ in the margin cannot be avoided. Consider the input distribution:

$$\begin{gathered}
p(X = x_1 = (\gamma, \sqrt{1-2\gamma^2}), Y = 1) = \frac{1}{2} - \epsilon \\
p(X = x_2 = (\gamma, -\sqrt{1-2\gamma^2}), Y = 1) = \epsilon \\
p(X = x_3 = (-\gamma, \sqrt{1-2\gamma^2}), Y = -1) = \frac{1}{2} - \epsilon \\
p(X = x_4 = (-\gamma, -\sqrt{1-2\gamma^2}), Y = -1) = \epsilon
\end{gathered}$$

It can be shown that the best margin that can be achieved on all four points by using the distance as a similarity is $2(\epsilon\gamma + 2\gamma^2)$.

All the results in this paper (and also the results of Balcan and Blum [1]) refer to consistent input distributions. Noisy input distributions, where some x might take either label with positive probability, are problematic when we use the definitions of Section 2: The weight $w(x)$ can depend only on x, but not on the label y, and so a positive weight yields a contribution from both labels. A point x with $\Pr(\,1|x\,)$ and $\Pr(\,-1|x\,)$ both high, cannot contribute much to the similarity-based classification margin (in the extreme case, if $\Pr(\,1|x\,) = \Pr(\,-1|x\,) = 0.5$, its contribution to the similarity-based classification margin will always be zero).

It is possible to use the results presented here also to obtain (rather messy) results for the noisy case by first removing examples with highly ambiguous labels, then applying Theorems 1 or 2, and finally correcting the weights to account for the negative contribution of the "wrong" label. The amount of this "correction", which will reduce the margin, can be bounded by the amount of allowed ambiguity, and the overall number of removed, highly ambiguous examples, can be bounded in terms of the error-rate. If the error-rate is bounded away from $\frac{1}{2}$, such an approach introduces only a multiplicative factor to both the resulting margin, and the associated margin-violations error-rate (note that in Theorem 1, for the consistent case, we only have an *additive* increase in the error-rate). However, since the hinge-loss on those examples that we removed might be extremely high, the deterioration of the hinge-loss guarantee is much worse. For this reason, a different approach might be appropriate.

We suggest changing the definition of the similarity-based classification margin, removing the effect of the label Y' and instead allowing both positive and negative weights in the range $[-1, +1]$, with the following as an alternative to the classification margin given in equation (1):

$$y\mathbf{E}_{X'}[w(X')K(x, X')]. \tag{28}$$

When the labels are balanced, i.e. $p(Y)$ is bounded away from 0, this yields strictly more flexible definitions, up to margin deterioration of $(\min_Y p(Y))$, for similarity goodness: the effect of the label can be incorporated into $w(x)$ by setting $w(x) \leftarrow w(x)\mathbf{E}_Y[Y/p(Y)|x](\min_Y p(Y))$. Nevertheless, all the learning results and methods of Balcan and Blum hold also using this revised definition of classification margin.

Under the revised definitions using (28), there is no problem handling noisy input distributions: Consider changing the weight mapping of equation (17) to

$$w(x) = \mathbf{E}[Y\alpha^*|x] \,/\, C. \tag{29}$$

We now no longer have to require that the label y is a deterministic function of x, and obtain the result of Theorems 1 and 2, with the same constants, for both consistent and noisy distributions, where the classification margin in equation (28) replaces that of equation (1) in Definitions 3 and 4. Note that the results do *not* depend on the label imbalance, and hold also when $p(y)$ is arbitrarily close to zero.

Acknowledgments

I would like to thank Avrim Blum and Nina Balcan for helpful discussions and quick responses and clarifications about their work; and Alexandre d'Aspremont for pointing me to [7]. It has been brought to my attention that the example of Section 5 was also independently constructed, and discussed in private communication, by Sanjoy Dasputa.

References

1. Balcan, M.F., Blum, A.: On a theory of learning with similarity functions. In: Proceedings of the 23rd International Conference on Machine Learning. (2006)
2. Jaakkola, T.S., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Advances in neural information processing systems 11. MIT Press (1999)
3. Viswanathan, S., Smola, A.J.: Fast kernels for string and tree matching. In: Advances in Neural Information Processing Systems 15. MIT Press (2003)
4. Smola, A.J., Schölkopf, B.: Learning with Kernels. MIT Press (2002)
5. McAllester, D.: Simplified pac-bayesian margin bounds. In: Proceedings of the 16th Conference on Computational Learning Theory. (2003)
6. Bartlett, P.L., Mendelson, S.: Rademacher and gaussian complexities: risk bounds and structural results. J. Mach. Learn. Res. **3** (2003) 463–482
7. Hettich, R., Kortanek, K.O.: Semi-infinite programming: theory, methods, and applications. SIAM Rev. **35** (1993) 380–429