Maximum Margin Multiclass Nearest Neighbors

Aryeh Kontorovich Roi Weiss

January 30, 2014

Abstract

We develop a general framework for margin-based multicategory classification in metric spaces. The basic work-horse is a margin-regularized version of the nearest-neighbor classifier. We prove generalization bounds that match the state of the art in sample size n and significantly improve the dependence on the number of classes k. Our point of departure is a nearly Bayes-optimal finite-sample risk bound independent of k. Although k-free, this bound is unregularized and non-adaptive, which motivates our main result: Rademacher and scale-sensitive margin bounds with a logarithmic dependence on k. As the best previous risk estimates in this setting were of order $\sqrt{k}$, our bound is exponentially sharper. From the algorithmic standpoint, in doubling metric spaces our classifier may be trained on n examples in $O(n^2 \log n)$ time and evaluated on new points in $O(\log n)$ time.

1 Introduction

Whereas the theory of supervised binary classification is by now fairly well developed, its multiclass extension continues to pose numerous novel statistical and computational challenges. On the algorithmic front, there is the basic question of how to adapt the hyperplane and kernel methods — ideally suited for two classes — to three or more. A host of new problems also arises on the statistical front. In the binary case, the VC-dimension characterizes the distribution-free sample complexity (Anthony & Bartlett, 1999) and tighter distribution-dependent bounds are available via Rademacher techniques (Bartlett & Mendelson, 2002; Koltchinskii & Panchenko, 2002). Characterizing the multiclass distribution-free sample complexity is far less straightforward, though impressive progress has been recently made (Daniely et al., 2011).

Following von Luxburg & Bousquet (2004); Gottlieb et al. (2010), we adopt a proximity-based approach to supervised multicategory classification in metric spaces. The principal motivation for this framework is two-fold:

(i) Many natural metrics, such as L_1, earthmover, and edit distance cannot be embedded in a Hilbert space without a large distortion (Enflo, 1969; Naor & Schechtman, 2007; Andoni & Krauthgamer, 2010). Any kernel

method is thus a priori at a disadvantage when learning to classify non-Hilbertian objects, since it cannot faithfully represent the data geometry.

(ii) Nearest neighbor-based classification sidesteps the issue of k-to-binary reductions — which, despite voluminous research, is still the subject of vigorous debate (Rifkin & Klautau, 2004; El-Yaniv et al., 2008). In terms of time complexity, the reductions approach faces an $\Omega(k)$ information-theoretic lower bound (Beygelzimer et al., 2009), while nearest neighbors admit solutions whose runtime does not depend on the number of classes.

Main results. Our contributions are both statistical and algorithmic in nature. On the statistical front, we open with the observation that the nearest-neighbor classifier's expected risk is at most twice the Bayes optimal plus a term that decays with sample size at a rate *not dependent* on the number of classes k (and continues to hold for $k = \infty$, Theorem 1). Although of interest as apparently the first "k-free" finite-sample result, it has the drawback of being *non-adaptive* in the sense of depending on properties of the unknown sampling distribution and failing to provide the learner with a usable data-dependent bound. This difficulty is overcome in our main technical contribution (Theorems 4 and 5), where we give a margin-based multiclass bound of order

$$\min\left\{\frac{1}{\gamma}\left(\frac{\log k}{n}\right)^{\frac{1}{D+1}}, \frac{1}{\gamma^{\frac{D}{2}}}\left(\frac{\log k}{n}\right)^{\frac{1}{2}}\right\}, \tag{1}$$

where k is the number of classes, n is sample size, D is the doubling dimension of the metric instance space and $0 < \gamma \leq 1$ is the margin. This matches the state of the art asymptotics in n for metric spaces and significantly improves the dependence on k, which hitherto was of order $\sqrt{k}$ (Zhang, 2002, 2004) or worse. The exponential dependence on some covering dimension (such as D) is in general inevitable, as shown by a standard no-free-lunch argument (Ben-David & Shalev-Shwartz, 2014), but whether (1) is optimal remains an open question.

On the algorithmic front, using the above bounds, we show how to efficiently perform Structural Risk Minimization (SRM) so as to avoid overfitting. This involves deciding *how many* and *which* sample points one is allowed to err on. We reduce this problem to minimal vertex cover, which admits a greedy 2-approximation. Our algorithm admits a significantly faster ε-approximate version in doubling spaces with a graceful degradation in ε of the generalization bounds, based on approximate nearest neighbor techniques developed by Gottlieb et al. (2010, 2013a). For a fixed doubling dimension and ε, our runtime is $O(n^2 \log n)$ for learning and $O(\log n)$ for evaluation on a test point. (Exact nearest neighbor requires $\Theta(n)$ evaluation time.) Finally, our generalization bounds and algorithm can be made adaptive to the intrinsic dimension of the data via a recent metric dimensionality-reduction technique (Gottlieb et al., 2013b).

Related work. Due to space constraints, we are only able to mention the most directly relevant results — and even these, not in full generality but rather with an eye to facilitating comparison to the present work. Supervised k-category classification approaches follow two basic paradigms: **(I)** defining a score function on point-label pairs and classifying by choosing the label with the optimal score and **(II)** reducing the problem to several binary classification problems. Regarding the second paradigm, the seminal paper of Allwein et al. (2001) unified the various error correcting output code (ECOC)-based multiclass-to-binary reductions under a single margin-based framework. Their generalization bound requires the base classifier to have VC-dimension $d_{\mathrm{VC}} < \infty$ (and hence does not apply to nearest neighbors or infinite-dimensional Hilbert spaces) and is of the form $\tilde{O}\Big(\frac{\log k}{\gamma}\sqrt{\frac{d_{\mathrm{VC}}}{n}}\Big)$. Langford & Beygelzimer (2005); Beygelzimer et al. (2009) gave k-free and $O(\log k)$ regret bounds, but these are conditional on the performance of the underlying binary classifiers as opposed to the unconditional bounds we provide in this paper.

As for the first paradigm, proximity is perhaps the most natural score function — and indeed, a formal analysis of the nearest neighbor classifier (Cover & Hart, 1967) much predated the first multiclass extensions of SVM (Weston & Watkins, 1999). Crammer & Singer (2002a,b) considerably reduced the computational complexity of the latter approach and gave a risk bound decaying as $\tilde{O}(k^2/n\gamma^2)$, for the separable case with margin γ. In an alternative approach based on choosing q prototype examples, Crammer et al. (2002) gave a risk bound with rate $\tilde{O}(q^{k/2}/\gamma\sqrt{n})$. Ben-David et al. (1995) characterized the PAC learnability of k-valued functions in terms of combinatorial dimensions, such as the Natarajan dimension d_{Nat}. Guermeur (2007, 2010) gave scale-sensitive analogues of these dimensions. He gave a risk bound decaying as $\tilde{O}\left(\frac{\log k}{\gamma}\sqrt{d_{\gamma\mathrm{Nat}}/n}\right)$, where $d_{\gamma\mathrm{Nat}}$ is a scale-sensitive Natarajan dimension — essentially replacing the finite VC dimension d_{VC} in Allwein et al. (2001) by $d_{\gamma\mathrm{Nat}}$. He further showed that for linear function classes in Hilbert spaces, $d_{\gamma\mathrm{Nat}}$ is bounded by $\tilde{O}(k^2/\gamma^2)$, resulting in a risk bound decaying as $\tilde{O}(k/\gamma^2\sqrt{n})$. To the best of our knowledge, the sharpest current estimate on the Natarajan dimension (for some special function classes) is $d_{\mathrm{Nat}} = \tilde{O}(k)$ with a matching lower bound of $\Omega(k)$ (Daniely et al., 2011). A margin-based Rademacher analysis of score functions (Mohri et al., 2012) yields a bound of order $\tilde{O}(k^2/\gamma\sqrt{n})$, and this is also the k-dependence obtained by Cortes et al. (2013) in a recent paper proposing a multiple kernel approach to multiclass learning. Closest in spirit to our work are the results of Zhang (2002, 2004), who used the chaining technique to achieve a Rademacher complexity with asymptotics $\tilde{O}\Big(\frac{1}{\gamma}\sqrt{\frac{k}{n}}\Big)$.

Besides the dichotomy of score functions vs. multiclass-to-binary reductions outlined above, multicategory risk bounds may also be grouped by the trichotomy of **(a)** combinatorial dimensions **(b)** Hilbert spaces **(c)** metric spaces (see Table 1). Category (a) is comprised of algorithm-independent results that give generalization bounds in terms of some combinatorial dimension of a fixed concept class (Allwein et al., 2001; Ben-David et al., 1995; Guermeur, 2007,

Paper	decay rate $\tilde{O}(\cdot)$	group
Allwein et al. (2001)‡	$\frac{\log k}{\gamma}\sqrt{\frac{d_{\mathrm{VC}}}{n}}$	(II,a)
Daniely et al. (2011)*†‡	$\frac{d_{\mathrm{Nat}}\log k}{n}$	(I,a)
Guermeur (2010)‡	$\frac{\log k}{\gamma}\sqrt{\frac{d_{\gamma\mathrm{Nat}}}{n}}$	(I,a)
Crammer & Singer (2002b)†	$\frac{k^2}{\gamma^2 n}$	(I,b)
Cortes et al. (2013)	$\frac{k^2}{\gamma\sqrt{n}}$	(I,b)
Guermeur (2010)	$\frac{k}{\gamma^2\sqrt{n}}$	(I,b)
Zhang (2004)	$\frac{1}{\gamma}\sqrt{\frac{k}{n}}$	(I,b)
current paper	$\frac{1}{\gamma^{D/2}}\sqrt{\frac{\log k}{n}}$	(I,c)
current paper	$\frac{1}{\gamma}\left(\frac{\log k}{n}\right)^{\frac{1}{1+D}}$	(I,c)

Table 1: Comparing various multiclass bounds. (*) Not margin-based. (†) Only for the separable case. (‡) Combinatorial dimension depends on k.

2010; Daniely et al., 2011). Multiclass extensions of SVM and related kernel methods (Weston & Watkins, 1999; Crammer & Singer, 2002a,b; Crammer et al., 2002; Cortes et al., 2013) fall into category (b). Category (c), consisting of agnostic[1] metric-space methods is the most sparsely populated. The pioneering asymptotic analysis of Cover & Hart (1967) was cast in a modern, finite-sample version by Ben-David & Shalev-Shwartz (2014), but only for binary classification. Unlike Hilbert spaces, which admit dimension-free margin bounds, we are not aware of any metric space risk bound that does not explicitly depend on some metric dimension D or covering numbers. The bounds in Ben-David & Shalev-Shwartz (2014); Gottlieb et al. (2013b) exhibit a characteristic "curse of dimensionality" decay rate of $O(n^{-1/(D+1)})$, but more optimistic asymptotics can be obtained (Guermeur, 2007, 2010; Zhang, 2002, 2004; Gottlieb et al., 2010). Although some sample lower bounds for proximity-based methods are known (Ben-David & Shalev-Shwartz, 2014), the optimal dependence on D and k is far from being fully understood.

2 Preliminaries

Metric Spaces. Given two metric spaces $(\mathcal{X}, d)$ and $(\mathcal{Z}, \rho)$, a function $f : \mathcal{X} \to \mathcal{Z}$ is called L-Lipschitz if $\rho(f(x), f(x')) \leq Ld(x, x')$ for all $x, x' \in \mathcal{X}$. (The real line $\mathbb{R}$ is always considered with its Euclidean metric $|\cdot|$.) The Lipschitz constant of f, denoted $\|f\|_{\mathrm{Lip}}$, is the smallest L for which f is L-Lipschitz. The distance between two sets $A, B \subset \mathcal{X}$ is defined by $d(A, B) = \inf_{x\in A, x'\in B} d(x, x')$. For a metric space $(\mathcal{X}, d)$, let λ be the smallest value such that every ball in

[1] in the sense of not requiring an a priori fixed concept class

$\mathcal{X}$ can be covered by λ balls of half the radius. The *doubling dimension* of $\mathcal{X}$ is $\mathrm{ddim}(\mathcal{X}) := \log_2 \lambda$. A metric is *doubling* when its doubling dimension is bounded. The ε-covering number of a metric space $(\mathcal{X}, d)$, denoted $\mathcal{N}(\varepsilon, \mathcal{X}, d)$, is defined as the smallest number of balls of radius ε that suffices to cover $\mathcal{X}$. It can be shown (e.g., Krauthgamer & Lee (2004)) that

$$\mathcal{N}(\varepsilon, \mathcal{X}, d) \leq \left(\frac{2\,\mathrm{diam}(\mathcal{X})}{\varepsilon}\right)^{\mathrm{ddim}(\mathcal{X})}, \tag{2}$$

where $\mathrm{diam}(\mathcal{X}) = \sup_{x,x' \in \mathcal{X}} d(x, x')$ is the diameter of $\mathcal{X}$.

The multiclass learning framework. Let $(\mathcal{X}, d)$ be a metric instance space with $\mathrm{diam}(\mathcal{X}) = 1$, $\mathrm{ddim}(\mathcal{X}) = D < \infty$, and $\mathcal{Y} \subseteq \mathbb{N}$ an at most countable label set. We observe a sample $S = (X_i, Y_i)_{i=1}^n \in \{\mathcal{X} \times \mathcal{Y}\}^n$ drawn iid from an unknown distribution P over $\mathcal{X} \times \mathcal{Y}$.

In line with paradigm (I) outlined in the Introduction, our classification procedure consists of optimizing a score function. In hindsight, the score at a test point will be determined by its labeled neighbors, but for now, we consider an unspecified collection $\mathcal{F}$ of functions mapping $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$. A score function $f \in \mathcal{F}$ induces the classifier $g_f : \mathcal{X} \to \mathcal{Y}$ via

$$g_f(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} f(x, y), \tag{3}$$

breaking ties arbitrarily. The *margin* of $f \in \mathcal{F}$ on (x, y) is defined by

$$\gamma_f(x, y) = \frac{1}{2}\left(f(x, y) - \sup_{y' \neq y} f(x, y')\right). \tag{4}$$

Note that g_f misclassifies (x, y) precisely when $\gamma_f(x, y) < 0$. One of our main objectives is to upper-bound the generalization error

$$\mathbb{P}(g_f(X) \neq Y) = \mathbb{E}[\mathbb{1}_{\{\gamma_f(X,Y)<0\}}].$$

To this end, we introduce two surrogate loss functions $\mathcal{L} : \mathbb{R} \to \mathbb{R}_+$:

$$\begin{aligned} \mathcal{L}_{\mathrm{cutoff}}(u) &= \mathbb{1}_{\{u<1\}} \\ \mathcal{L}_{\mathrm{margin}}(u) &= \mathrm{T}_{[0,1]}(1-u), \end{aligned}$$

where

$$\mathrm{T}_{[a,b]}(z) = \max\{a, \min\{b, z\}\} \tag{5}$$

is the truncation operator. The empirical loss $\widehat{\mathbb{E}}[\mathcal{L}(\gamma_f)]$ induced by any of the loss functions above is $\frac{1}{n}\sum_{i=1}^n \mathcal{L}(\gamma_f(X_i, Y_i))$. All probabilities $\mathbb{P}(\cdot)$ and expectations $\mathbb{E}[\cdot]$ are with respect to the sampling distribution P. We will write $\mathbb{E}_S$ to indicate expectation over a sample (i.e., over P^n).

3 Risk bounds

In this section we analyze the statistical properties of nearest-neighbor multicategory classifiers in metric spaces. In Section 3.1, Theorem 1, we record the observation that the 1-nearest neighbor classifier is nearly Bayes optimal, with a risk decay that does not depend on the number of classes k. Of course, the 1-naive nearest neighbor is well-known to overfit. This is reflected in the non-adaptive nature of the analysis: the bound is stated in terms of properties of the unknown sampling distribution, and fails to provide the learner with a usable data-dependent bound.

To achieve the latter goal, we develop a margin analysis in Section 3.2. Our main technical result is Lemma 2, from which the logarithmic dependence on k claimed in (1) follows. Although not k-free like the Bayes excess risk bound of Theorem 1, $O(\log k)$ is exponentially sharper than the current state of the art (Zhang, 2002, 2004). Whether a k-free metric entropy bound is possible is currently left as an open problem.

The metric entropy bound of Lemma 2 facilitates two approaches to bounding the risk: via Rademacher complexity (Section 3.2.2) and via scale-sensitive techniques in the spirit of Guermeur (2007) (Section 3.2.3). In Section 3.2.4 we combine these two margin bounds by taking their minimum. The resulting bound will be used in Section 4 to perform efficient Structural Risk Minimization.

3.1 Multiclass Bayes near-optimality

In this section, $(\mathcal{X}, d)$ is a metric space and $\mathcal{Y}$ is an at most countable (possibly infinite) label set. A sample $S = (X_i, Y_i)_{i=1}^n$ is drawn iid from an unknown distribution P over $\mathcal{X} \times \mathcal{Y}$. For $x \in \mathcal{X}$ let $(X_{\pi_1(x)}, Y_{\pi_1(x)})$ be its nearest neighbor in S:

$$\pi_1(x) = \operatorname*{argmin}_{i \in [n]} d(X_i, x).$$

Thus, the nearest-neighbor classifier g_{NN} is given by

$$g_{\text{NN}}(x) = Y_{\pi_1(x)}. \tag{6}$$

Define the function $\boldsymbol{\eta} : \mathcal{X} \to \mathbb{R}^{\mathcal{Y}}$ by

$$\boldsymbol{\eta}(x) = \mathbb{P}(Y = \cdot \,|\, X = x).$$

The *Bayes optimal* classifier g^* — i.e., one that minimizes $\mathbb{P}(g(X) \neq Y)$ over all measurable $g \in \mathcal{Y}^{\mathcal{X}}$ — is well-known to have the form

$$g^*(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \eta_y(x),$$

where ties are broken arbitrarily. Our only distributional assumption is that $\boldsymbol{\eta}$ is L-Lipschitz with respect to the sup-norm. Namely, for all $x, x' \in \mathcal{X}$, we have

$$\|\boldsymbol{\eta}(x) - \boldsymbol{\eta}(x')\|_\infty \equiv \sup_{y \in \mathcal{Y}} |\eta_y(x) - \eta_y(x')| \leq L d(x, x').$$

This is a direct analogue of the Lipschitz assumption for the binary case (Cover & Hart, 1967; Ben-David & Shalev-Shwartz, 2014). We make the additional standard assumption that $\mathcal{X}$ has a finite doubling dimension: $\mathrm{ddim}(\mathcal{X}) = D < \infty$. The Lipschitz and doubling assumptions are sufficient to extend the finite-sample analysis of binary nearest neighbors (Ben-David & Shalev-Shwartz, 2014) to the multiclass case:

Theorem 1.

$$\mathbb{E}_S\left[\mathbb{P}(g_{\text{NN}}(X) \neq Y)\right] \leq 2\mathbb{P}(g^*(X) \neq Y) + \frac{4L}{n^{1/(D+1)}}.$$

Note that the bound is independent of the number of classes k and holds even for $k = \infty$. The proof is deferred to Appendix A.

3.2 Multiclass margin bounds

Here again $(\mathcal{X}, d)$ is a metric space, but now the label set $\mathcal{Y}$ is assumed finite: $|\mathcal{Y}| = k < \infty$. As before, $S = (X_i, Y_i)_{i=1}^n$ with $(X_i, Y_i) \sim P$ iid. It will be convenient to write $S^y = \{X_i : Y_i = y, i \in [n]\}$ for the subset of examples with label y. The metric induces the natural score function $f_{\text{NN}}(x, y) = -d(x, S^y)$ with corresponding nearest-neighbor classifier

$$g_{\text{NN}}(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} f_{\text{NN}}(x, y), \tag{7}$$

easily seen to be identical to the one in (6). At this point we make the simple but crucial observation that the function $f_{\text{NN}}(\cdot, y) : \mathcal{X} \to \mathbb{R}$ is 1-Lipschitz. This will enable us to generalize the powerful Lipschitz extension framework of von Luxburg & Bousquet (2004) to $|\mathcal{Y}| > 2$.

We will need a few definitions. Let F_L be the collection of all L-Lipschitz functions from $\mathcal{X}$ to $\mathbb{R}$ and put $\mathcal{F}_L = F_L \times \mathcal{Y}$. Since each $f \in \mathcal{F}_L$ maps $\mathcal{X} \times \mathcal{Y}$ to $\mathbb{R}$, the margin $\gamma_f(x, y)$ is well-defined via (4). Putting

$$\begin{aligned} y_f^*(x) &= \operatorname*{argmax}_{y \in \mathcal{Y}} f(x, y), \\ \gamma_f^*(x) &= \gamma_f(x, y_f^*(x)), \end{aligned}$$

we define the *projection* Φ_f:

$$\Phi_f(x, y) = \begin{cases} \gamma_f^*(x), & \text{if } y = y_f^*(x) \\ -\gamma_f^*(x), & \text{otherwise.} \end{cases}$$

Finally, we define $\mathcal{H}_L$ as the truncated (as in (5)) projections of functions in $\mathcal{F}_L$:

$$\mathcal{H}_L = \left\{(x, y) \mapsto \mathrm{T}_{[-1,1]}\left(\Phi_f(x, y)\right) : f \in \mathcal{F}_L\right\}. \tag{8}$$

Thus, $\mathcal{H}_L$ is the set of functions $h_f : \mathcal{X} \times \mathcal{Y} \to [-1, 1]$, where each $h_f(\cdot, y)$ is L-Lipschitz and $h_f(x, y) = \pm \mathrm{T}_{[-1,1]}(\gamma_f^*(x))$, depending upon whether $y = y_f^*(x)$, see Figure 1 (left).

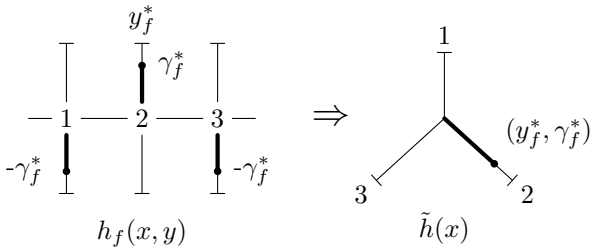

Figure 1: The mapping in Lemma 2 with $|\mathcal{Y}| = 3$.

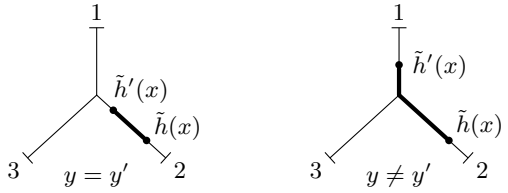

Figure 2: The metric $\rho(\tilde{h}(x), \tilde{h}'(x))$ with $|\mathcal{Y}| = 3$.

3.2.1 Bounding the metric entropy

Our main technical result is a bound on the metric entropy of $\mathcal{H}_L$, which will be used to obtain error bounds (Theorems 4 and 5) for classifiers derived from this function class. The analysis differs from previous bounds (see Table 1) by explicitly taking advantage of the mutual exclusive nature of the labels, obtaining an exponential improvement in terms of the number of classes k. Endow $\mathcal{H}_L$ with the sup-norm

$$\|\cdot\|_\infty = \sup_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} |\ \cdot\ | .$$

Lemma 2. *For any $\varepsilon > 0$,*

$$\log \mathcal{N}(\varepsilon, \mathcal{H}_L, \|\cdot\|_\infty) \leq \left(\frac{16L}{\varepsilon}\right)^D \log\left(\frac{5k}{\varepsilon}\right).$$

Proof. By the definition of $\mathcal{H}_L$, for all $h_f \in \mathcal{H}_L$ and $x \in \mathcal{X}$ there is at most one $y \in \mathcal{Y}$ such that $h_f(x, y) > 0$. In addition, if $h_f(x, y) = c > 0$, then $h_f(x, y') = -c$ for all $y' \neq y$. Since $\gamma_f^*(x) \geq 0$, we may reparametrize $h_f(x, y)$ by $(y_f^*(x), \gamma_f^*(x)) \in \mathcal{Y} \times [0, 1]$, see Figure 1. To complete the mapping $h_f \mapsto (y_f^*, \gamma_f^*)$, define the following star-like metric ρ over $\mathcal{Y} \times [0, 1]$ (see Figure 2):

$$\rho((y,\gamma),(y',\gamma')) = \begin{cases} |\gamma - \gamma'| & y = y' \\ \gamma + \gamma' & y \neq y' \end{cases} .$$

Let $\tilde{H}_L$ be the collection of functions $\tilde{h} : \mathcal{X} \to \mathcal{Y} \times [0, 1]$ that are L-Lipschitz:

$$\rho(\tilde{h}(x), \tilde{h}(x')) \leq L d(x, x'), \qquad x, x' \in \mathcal{X}.$$

It is easily verified that the metric space $(\mathcal{H}_L, \|\cdot\|_\infty)$ is isometric to $(\tilde{H}_L, \rho_\infty)$ with

$$\rho_\infty(\tilde{h}, \tilde{h}') = \sup_{x \in \mathcal{X}} \rho(\tilde{h}(x), \tilde{h}'(x)).$$

Thus, $\mathcal{N}(\varepsilon, \mathcal{H}_L, \|\cdot\|_\infty) = \mathcal{N}(\varepsilon, \tilde{H}_L, \rho_\infty)$, and we proceed to bound the latter.[2] Fix a covering of $\mathcal{X}$ consisting of $|N| = \mathcal{N}(\varepsilon/8L, \mathcal{X}, d)$ balls $\{U_1, \ldots, U_{|N|}\}$ of radius $\varepsilon' = \varepsilon/8L$ and choose $|N|$ points $N = \{x_i \in U_i\}_{i=1}^{|N|}$. Construct $\widehat{H} \subset \tilde{H}_{2L}$ as follows. At every point $x_i \in N$ select one of the classes $y \in \mathcal{Y}$ and set $\widehat{h}(x_i) = (y, \gamma(x_i))$ with $\gamma(x_i)$ some multiple of $2L\varepsilon' = \varepsilon/4$, while maintaining $\|\widehat{h}\|_{\text{Lip}} \leq 2L$. Construct a $2L$-Lipschitz extension for $\widehat{h}$ from N to all over $\mathcal{X}$ (such an extension always exists, (McShane, 1934; Whitney, 1934)). We claim that every classifier in $\mathcal{H}_L$, via its twin $\tilde{h} \in \tilde{H}_L$, is close to some $\widehat{h} \in \widehat{H}$, in the sense that $\rho_\infty(\tilde{h}, \widehat{h}) \leq \varepsilon$. Indeed, every point $x \in \mathcal{X}$ is $2\varepsilon'$-close to some point $x_i \in N$, and since $\tilde{h}$ is L-Lipschitz and $\widehat{h}$ is $2L$-Lipschitz,

$$\begin{aligned}
\rho(\tilde{h}(x), \widehat{h}(x)) &\leq \rho(\tilde{h}(x), \tilde{h}(x_i)) \\
&\quad + \rho(\tilde{h}(x_i), \widehat{h}(x_i)) \\
&\quad + \rho(\widehat{h}(x_i), \widehat{h}(x)) \\
&\leq L d(x, x_i) + \varepsilon/4 + 2L d(x, x_i) \\
&\leq \varepsilon.
\end{aligned}$$

Thus, $\widehat{H}$ provides an ε-cover for $\tilde{H}_L$ (and hence for $\mathcal{H}_L$). Note that $|\widehat{H}| \leq (\lceil 4k/\varepsilon \rceil + 1)^{|N|}$, since by construction, functions $\widehat{h}$ are determined by their values on N, which at a given point can take one of $\lceil 4k/\varepsilon \rceil + 1$ possible values. Since by (2) we have $|N| = \mathcal{N}(\varepsilon/8L, \mathcal{X}, d) \leq \left(\frac{16L}{\varepsilon}\right)^D$ the bound follows. □

A tighter bound is possible when the metric space $(\mathcal{X}, d)$ possesses two additional properties:

1. $(\mathcal{X}, d)$ is *connected* if for all $x, x' \in \mathcal{X}$ and all $\varepsilon > 0$, there is a finite sequence of points $x = x_1, x_2, \ldots, x_m = x'$ such that $d(x_i, x_{i+1}) < \varepsilon$ for all $1 \leq i < m$.

2. $(\mathcal{X}, d)$ is *centered* if for all $r > 0$ and all $A \subset \mathcal{X}$ with $\text{diam}(A) \leq 2r$, there exists a point $x \in \mathcal{X}$ such that $d(x, a) \leq r$ for all $a \in A$.

Lemma 3. *If $(\mathcal{X}, d)$ is connected and centered, then*

$$\log \mathcal{N}(\varepsilon, \mathcal{H}_L, \|\cdot\|_\infty) = O\left(\left(\frac{L}{\varepsilon}\right)^D \log k + \log\left(\frac{1}{\varepsilon}\right)\right).$$

Proof. With the additional assumptions on $\mathcal{X}$ we follow the proof idea in Kolmogorov & Tikhomirov (1959) and demonstrate the tighter bound $|\widehat{H}| \leq (\lceil 4k/\varepsilon \rceil + 1)(2k+1)^{|N|-1} =$

[2]The remainder of the proof is based on a technique communicated to us by R. Krauthgamer, a variant of the classic Kolmogorov & Tikhomirov (1959) method.

$O((2k)^{|N|}/\varepsilon)$. Here $\widehat{H}$ is constructed as in the proof for Lemma 2 but now each $x_i \in N$ is taken to be a "center" of U_i, as furnished by Property 2 above. Let $x_j \in N$. Since $\mathcal{X}$ is connected, we may traverse a path from x_1 to x_j via the cover points $x_1 = x_{i_1}, x_{i_2}, \ldots, x_{i_m} = x_j$, such that the distance between any two successive points $(x_{i_l}, x_{i_{l+1}})$ is at most $2\varepsilon' = \varepsilon/4L$. Since $\widehat{h}$ is $2L$-Lipschitz, on any two such points the value of $\widehat{h}$ can change by at most $\varepsilon/2$. Thus, given the value $\widehat{h}(x_{i_l})$, the value of $\widehat{h}(x_{i_{l+1}})$ can take one of at most $2k+1$ values (as Figure 2 shows, at the star's hub, $\widehat{h}(x_{i_{l+1}})$ can take one of $2k+1$ values, while at one of the spokes only 5 values are possible). So we are left to choose the value of $\widehat{h}$ on the point x_1 to be one from the $\lceil 4k/\varepsilon \rceil + 1$ possible values. The bounds on $|\widehat{H}|$ and the metric entropy follow. □

3.2.2 Rademacher analysis

The *Rademacher complexity* of the set of functions $\mathcal{H}_L$ is defined by

$$\mathcal{R}_n(\mathcal{H}_L) = \mathbb{E}\left[\sup_{h \in \mathcal{H}_L} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(X_i, Y_i)\right], \tag{9}$$

where the σ_i are n independent random variables with $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = 1/2$. In Appendix B, we invoke Lemma 2 to derive the bound

$$\mathcal{R}_n(\mathcal{H}_L) \leq 2L \left(\frac{\log 5k}{n}\right)^{1/(D+1)}, \tag{10}$$

which in turn implies "half" of our main risk estimate (1):

Theorem 4. *With probability at least $1-\delta$, for all $L > 0$ and every $f \in \mathcal{F}_L$ with its projected version $h_f \in \mathcal{H}_L$,*

$$\mathbb{P}(g_f(X) \neq Y) \leq \widehat{\mathbb{E}}\left[\mathcal{L}(h_f)\right] + \Delta_{\text{Rad}}(n, L, \delta),$$

where g_f is the classifier defined in (3), $\mathcal{L}$ is any of the loss functions defined in Section 2 and $\Delta_{\text{Rad}}(n, L, \delta)$ is at most

$$8L \left(\frac{\log 5k}{n}\right)^{\frac{1}{D+1}} + \sqrt{\left(\frac{\log \log_2 2L}{n}\right)_+} + \sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

3.2.3 Scale-sensitive analysis

The following Theorem , proved in Appendix C, is an adaptation of Guermeur (2007, Theorem 1), using Lemma 2.

Theorem 5. *With probability at least $1-\delta$, for all $L > 0$ and every $f \in \mathcal{F}_L$ with its induced $h_f \in \mathcal{H}_L$,*

$$\mathbb{P}(g_f(X) \neq Y) \leq \widehat{\mathbb{E}}\left[\mathcal{L}_{\text{cutoff}}(h_f)\right] + \Delta_{\text{fat}}(n, L, \delta),$$

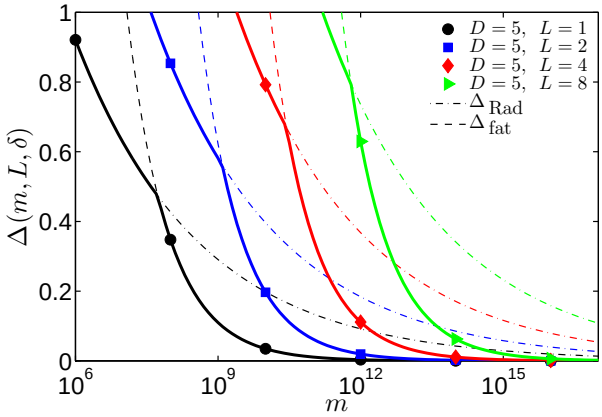

Figure 3: The combined complexity bounds ($k = 10, \delta = 0.01$).

where $\Delta_{\mathrm{fat}}(n, L, \delta)$ *is at most*

$$\sqrt{\frac{2}{n}\left(2\left(16L\right)^{D}\log\left(20k\right) + \ln\left(\frac{2L}{\delta}\right)\right)} + \frac{1}{n}.$$

3.2.4 Combined Bound

Taking $\mathcal{L} = \mathcal{L}_{\mathrm{cutoff}}$ in Theorem 4 we can merge the above two bounds by taking their minimum. Namely, Theorem 5 holds with $\Delta(n, L, \delta) = \min\left\{\Delta_{\mathrm{Rad}}(n, L, \delta), \Delta_{\mathrm{fat}}(n, L, \delta)\right\}$ in place of $\Delta_{\mathrm{fat}}(n, L, \delta)$, see Figure 3. The resulting risk decay rate is of order

$$\min\left\{L\left(\frac{\log k}{n}\right)^{\frac{1}{D+1}}, L^{\frac{D}{2}}\left(\frac{\log k}{n}\right)^{\frac{1}{2}}\right\},$$

as claimed in (1). In terms of the number of classes k, our bound compares favorably to those in Allwein et al. (2001); Guermeur (2007, 2010), and more recently in Daniely et al. (2011), which have a k-dependence of $O(d_{\mathrm{Nat}} \log k)$, where d_{Nat} is the (scale-sensitive, k-dependent) Natarajan dimension of the multiclass hypothesis class. The optimal dependence of the risk on k is an intriguing open problem.

4 Algorithm

Theorems 4 and 5 yield generalization bounds of the schematic form

$$\mathbb{P}(g(X) \neq Y) \leq \widehat{\mathbb{E}}[\mathcal{L}] + \Delta(n, L, \delta). \tag{11}$$

The free parameter L in (11) controls (roughly speaking) the bias-variance tradeoff: for larger L, we may achieve a smaller empirical loss $\widehat{\mathbb{E}}[\mathcal{L}]$ at the expense

of a larger hypothesis complexity $\Delta(n, L, \delta)$. Our Structural Risk Minimization (SRM) consists of seeking the optimal L — i.e., one that minimizes the right-hand side of (11) — via the following high-level procedure:

1. For each $L > 0$, minimize $\widehat{\mathbb{E}}[\mathcal{L}(h_f)]$ over $f \in \mathcal{F}_L$.
2. Choose the optimal L^* and its corresponding classifier g_f with $f \in \mathcal{F}_{L^*}$.

Minimizing the empirical loss. Let $S = (X_i, Y_i)_{i=1}^n$ be the training sample and $L > 0$ a given maximal allowed Lipschiz constant. We will say that a function $h \in \mathcal{H}_L$ is *inconsistent* with a sample point (x, y) if $h(x, y) < 1$ (i.e., if the margin of h on (x, y) is less than one). Denote by $\widehat{m}(L)$ the smallest possible number of sample points on which a function $h \in \mathcal{H}_L$ may be inconsistent:

$$\widehat{m}(L) = \min_{h \in \mathcal{H}_L} \widehat{\mathbb{E}}[\mathcal{L}_{\text{cutoff}}(h)].$$

Thus, our SRM problem consists of finding

$$L^* = \operatorname*{argmin}_{L>0} \left\{ \widehat{m}(L) + \Delta(n, L, \delta) \right\}.$$

For $k = 2$, Gottlieb et al. (2010) reduced the problem of computing $\widehat{m}(L)$ to one of finding a minimal vertex cover in a bipartite graph (by König's theorem, the latter is efficiently computable as a maximal matching). We will extend this technique to $k > 2$ as follows. Define the k-partite graph $G_L = (\{V^y\}_{y=1}^k, E)$, where each vertex set V^y corresponds to the sample points S^y with label y. Now in order for $h \in \mathcal{H}_L$ to be consistent with the points (X_i, Y_i) and (X_j, Y_j) for $Y_i \neq Y_j$, the following relation must hold:

$$Ld(X_i, X_j) \geq 2. \tag{12}$$

Hence, we define the edges of G_L to consist of all point pairs violating (12):

$$(X_i, X_j) \in E \quad \Longleftrightarrow \quad (Y_i \neq Y_j) \wedge (d(X_i, X_j) < 2/L).$$

Since removing either of X_i, X_j in (12) also deletes the violating edge, $\widehat{m}(L)$ is by construction equivalent to the size of the minimum vertex cover for G_L. Although minimum vertex cover is NP-hard to compute (and even hard to approximate within a factor of 1.3606, (Dinur & Safra, 2005)), a 2-approximation may be found in $O(n^2)$ time (Papadimitriou & Steiglitz, 1998). This yields a 2-approximation $\tilde{m}(L)$ for $\widehat{m}(L)$.

Optimizing over L. Equipped with an efficient routine for computing $\tilde{m}(L) \leq 2\widehat{m}(L)$, we now seek an $L > 0$ that minimizes

$$Q(L) := \tilde{m}(L) + \Delta(n, L, \delta). \tag{13}$$

Since the Lipschitz constant induced by the data is determined by the $\binom{n}{2}$ distances among the sample points, we need only consider $O(n^2)$ values of L.

Rather a brute-force searching all of these values, Theorem 7 of Gottlieb et al. (2010) shows that using an $O(\log n)$ time binary search over the values of L, one may approximately minimize $Q(L)$, which in turn yields an approximate solution to (11). The resulting procedure has runtime $O(n^2 \log n)$ and guarantees an $\tilde{L}$ for which

$$Q(\tilde{L}) \quad \leq \quad 4\left[\widehat{m}(L^*) + \Delta(n, L^*, \delta)\right]. \tag{14}$$

Classifying test points. Given the nearly optimal Lipschitz constant $\tilde{L}$ computed above we construct the approximate (within a factor of 4) empirical risk minimizer $h^* \in \mathcal{H}_{\tilde{L}}$. The latter partitions the sample into $S = S_0 \cup S_1$, where S_1 consists of the points on which h^* is consistent and $S_0 = S \setminus S_1$. Evaluating h^* on a test point amounts to finding its nearest neighbor in S_1. Although in general metric spaces, nearest-neighbors search requires $\Omega(n)$ time, for doubling spaces, an exponential speedup is available via approximate nearest neighbors (see Section 5).

5 Extensions

In this section, we discuss two approaches that render the methods presented above considerably more efficient in terms of runtime and generalization bounds. The first is based on the fact that in doubling spaces, hypothesis evaluation time may be reduced from $O(n)$ to $O(\log n)$ at the expense of a very slight degradation of the generalization bounds. The second relies on a recent metric dimensionality reduction result. When the data is "close" to being $\tilde{D}$-dimensional, with $\tilde{D}$ much smaller than the ambient metric space dimension D, both the evaluation runtime and the generalization bounds may be significantly improved — depending essentially on $\tilde{D}$ rather than D.

5.1 Exponential speedup via approximate NN

If $(\mathcal{X}, d)$ is a metric space and $x^* \in E \subset \mathcal{X}$ is a minimizer of $d(x, x')$ over $x' \in E$, then x^* is a *nearest neighbor* of x in E. A simple information-theoretic argument shows that the time complexity of computing an exact nearest neighbor in general metric spaces has $\Omega(n)$ time complexity. However, an exponential speedup is possible if (i) $\mathcal{X}$ is a doubling space and (ii) one is willing to settle for approximate nearest neighbors. A $(1+\eta)$ nearest neighbor oracle returns an $\tilde{x} \in E$ such that

$$d(x, x^*) \leq d(x, \tilde{x}) \leq (1+\eta) d(x, x^*). \tag{15}$$

We will use the fact that in a doubling space, one may precompute a $(1+\eta)$ nearest neighbor data structure in $(2^{O(\mathrm{ddim}(\mathcal{X}))} \log n + \eta^{-O(\mathrm{ddim}(\mathcal{X}))})n$ time and evaluate it on a test point in $2^{O(\mathrm{ddim}(\mathcal{X}))} \log n + \eta^{-O(\mathrm{ddim}(\mathcal{X}))}$ time Cole & Gottlieb (2006); Har-Peled & Mendel (2006). The approximate nearest neighbor oracle induces an η-approximate version of g_{NN} in defined (7). After performing SRM

as described in Section 4, we are left with a subset $S_1 \subset S$ of the sample, which will be used to label test points. More precisely, the predicted label of a test point will be determined by its η-nearest neighbor in S_1.

The exponential speedup afforded by approximate nearest neighbors comes at the expense of mildly degraded generalization guarantees. The modified generalization bounds are derived in three steps, whose details are deferred to Appendix D:
(i) We cast the evaluation of $h \in \mathcal{H}_L$ in (8) as a nearest neighbor calculation with a corresponding $\tilde{h}$ induced by the $(1+\eta)$ approximate nearest neighbor oracle. The nearest-neighbor formulation of h is essentially the one obtained by von Luxburg & Bousquet (2004):

$$h(x,y) = \frac{1}{2}\left(\min_{S_1}\{\xi(y,y') + Ld(x,x')\} + \max_{S_1}\{\xi(y,y') - Ld(x,x')\}\right), \tag{16}$$

where $(x',y') \in S_1$ and $\xi(y,y') = 2\mathbb{1}_{\{y=y'\}} - 1$.
(ii) We observe a simple relation between h and $\tilde{h}$:

$$\|h - \tilde{h}\|_\infty \equiv \sup_{x\in\mathcal{X}, y\in\mathcal{Y}} |h(x,y) - \tilde{h}(x,y)| \le 2\eta.$$

(iii) Defining the 2η-perturbed function class

$$\mathcal{H}_{L,2\eta} = \{\mathrm{T}_{[-1,1]}(h') : \|h' - h\|_\infty \le 2\eta, h \in \mathcal{H}_L\},$$

we relate its metric entropy to that of $\mathcal{H}_L$:

Lemma 6. *For $\varepsilon > 2\eta > 0$, we have*

$$\mathcal{N}(\varepsilon, \mathcal{H}_{L,2\eta}, \|\cdot\|_\infty) \quad \le \quad \mathcal{N}(\varepsilon - 2\eta, \mathcal{H}_L, \|\cdot\|_\infty).$$

The metric entropy estimate for $\mathcal{H}_{L,2\eta}$ readily yields η-perturbed versions of Theorems 4 and 5. From the standpoint of generalization bounds, the effect of the η-perturbation on $\mathcal{H}_L$ amounts, roughly speaking, to replacing L by $L(1+O(\eta))$, which constitutes a rather benign degradation.

5.2 Adaptive dimensionality reduction

The generalization bound in (1) and the runtime of our sped-up algorithm in Section 5.1 both depend exponentially on the doubling dimension of the metric space. Hence, even a modest dimensionality reduction could lead to dramatic savings in algorithmic and sample complexities. The standard Euclidean dimensionality-reduction tool, PCA, until recently had no metric analogue — at least not with rigorous performance guarantees. The technique proposed in Gottlieb et al. (2013b) may roughly be described as a metric analogue of PCA.

A set $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$ inherits the metric d of $\mathcal{X}$ and hence $\mathrm{ddim}(X) \leq \mathrm{ddim}(\mathcal{X})$ is well-defined. We say that $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_n\} \subset \mathcal{X}$ is an (α, β)-*perturbation* of X if $\sum_{i=1}^n d(x_i, \tilde{x}_i) \leq \alpha$ and $\mathrm{ddim}(\tilde{X}) \leq \beta$. Intuitively, the data is "essentially" low-dimensional if it admits an (α, β)-perturbation with small α, β, which leads to improved Rademacher estimates. The *empirical Rademacher complexity* of $\mathcal{H}_L$ on a sample $S = (X, Y) \in \mathcal{X}^n \times \mathcal{Y}^n$ is given by

$$\widehat{\mathcal{R}}_n(\mathcal{H}_L; S) = \mathbb{E}\left[\sup_{h \in \mathcal{H}_L} \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i, Y_i) \,\middle|\, S\right]$$

and is related to $\mathcal{R}_n$ defined in (9) via

$$\begin{aligned} \mathcal{R}_n(\mathcal{H}_L) &= \mathbb{E}_S\left[\widehat{\mathcal{R}}_n(\mathcal{H}_L; S)\right] \\ \mathbb{P}\left(\left|\mathcal{R}_n - \widehat{\mathcal{R}}_n\right| \geq \varepsilon\right) &\leq 2\exp(-\varepsilon^2 n/2), \end{aligned}$$

where the identity is obvious and the inequality is a simple consequence of measure concentration (Mohri et al., 2012). Hence, up to small changes in constants, the two may be used in generalization bounds such as Theorem 4 interchangeably. The data-dependent nature of $\widehat{\mathcal{R}}_n$ lets us exploit essentially low-dimensional data (see Appendix E):

Theorem 7. *Let $S = (X, Y) \in \mathcal{X}^n \times \mathcal{Y}^n$ be the training sample and suppose that X admits an (α, β)-perturbation $\tilde{X}$. Then*

$$\widehat{\mathcal{R}}_n(\mathcal{H}_L; S) = O\left(L\left(\alpha + \left(\frac{\log k}{n}\right)^{\frac{1}{1+\beta}}\right)\right). \tag{17}$$

A pleasant feature of the bound above is that it does not depend on $\mathrm{ddim}(\mathcal{X})$ (the dimension of the ambient space) or even on $\mathrm{ddim}(X)$ (the dimension of the data). Note the inherent tradeoff between the distortion α and dimension β, with some non-trivial (α^*, β^*) minimizing the right-hand side of (17). Although computing the optimal (α^*, β^*) seems computationally difficult, Gottlieb et al. (2013b) were able to obtain an efficient $(O(1), O(1))$-*bicriteria approximation*. Namely, their algorithm computes an $\tilde{\alpha} \leq c_0\alpha^*$ and $\tilde{\beta} \leq c_1\beta^*$, with the corresponding perturbed set $\tilde{X}$, for universal constants c_0, c_1, with a runtime of $2^{O(\mathrm{ddim}(X))} n \log n + O(n \log^5 n)$.

The optimization routine over (α, β) may then be embedded inside our SRM optimization over the Lipschitz constant L in Section 4. The end result will be a nearly optimal (in the sense of (14)) Lipschitz constant $\tilde{L}$, which induces the partition $S = S_0 \cup S_1$, as well as $(\tilde{\alpha}, \tilde{\beta})$, which induce the perturbed set $\tilde{S}_1$. To evaluate our hypothesis on a test point, we may invoke the $(1+\eta)$-approximate nearest-neighbor routine from Section 5.1. This involves a precomputation of time complexity $(2^{O(\tilde{\beta})} \log n + \eta^{-O(\tilde{\beta})})n$, after which new points are classified in $2^{O(\tilde{\beta})} \log n + \eta^{-O(\tilde{\beta})}$ time. Note that the evaluation time complexity depends only on the "intrinsic dimension" $\tilde{\beta}$ of the data, rather than the ambient metric space dimension.

References

Allwein, Erin L, Schapire, Robert E, and Singer, Yoram. Reducing multiclass to binary: A unifying approach for margin classifiers. *JMLR*, 1:113–141, 2001.

Andoni, A. and Krauthgamer, R. The computational hardness of estimating edit distance. *SICOMP*, 39(6):2398-2429, 2010.

Anthony, M. and Bartlett, P. *Neural network learning: theoretical foundations*. Cambridge University Press, 1999.

Bartlett, P. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463-482, 2002.

Ben-David, S. and Shalev-Shwartz, S. Understanding machine learning: From theory to algorithms. Cambridge University Press, 2014.

Ben-David, S., Cesa-Bianchi, N., Haussler, D., and Long, P. Characterizations of learnability for classes of $\{0, \cdots, n\}$-valued functions. *J. Comput. System Sci.*, 50(1):74–86, 1995.

Beygelzimer, A., Langford, J., and Ravikumar, P. Error-correcting tournaments. *ALT*, 2009.

Cole, R. and Gottlieb, L. Searching dynamic point sets in spaces with bounded doubling dimension. *STOC*, 2006.

Cortes, C., Mohri, M., and Rostamizadeh, A. Multi-class classification with maximum margin multiple kernel. ICML, 2013.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Info. Theo.,* , 13(1):21-27, 1967.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265-292, 2002a.

Crammer, K. and Singer, Y. On the learnability and design of output codes for multiclass problems. *Mach. Learn.*, 47(2-3):201-233, 2002b.

Crammer, K., Gilad-Bachrach, R., Navot, A., and Tishby, N. Margin analysis of the lvq algorithm. *NIPS*, 2002.

Daniely, A., Sabato, S., Ben-David, S., and Shalev-Shwartz, S. Multiclass learnability and the erm principle. *JMLR - Proceedings Track*, 19:207-232, 2011.

Dinur, I. and Safra, S. On the hardness of approximating minimum vertex cover. *Ann. Math.*, 162(1):439-485, 2005.

Dudley, R.M. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *J. Func. Anal.*, 1(3):290-330, 1967.

El-Yaniv, R., Pechyony, D., and Yom-Tov, E. Better multiclass classification via a margin-optimized single binary problem. *Patt. Rec. Lett.*, 29(14):1954-1959, 2008.

Enflo, P. On the nonexistence of uniform homeomorphisms between L_p-spaces. *Ark. Mat.*, 8:103-105, 1969.

Gottlieb, L., Kontorovich, A., and Krauthgamer, R. Efficient classification for metric data. *COLT*, 2010.

Gottlieb, L., Kontorovich, A., and Krauthgamer, R. Efficient regression in metric spaces via approximate lipschitz extension. *SIMBAD*, 2013a.

Gottlieb, L., Kontorovich, A., and Krauthgamer, R. Adaptive metric dimensionality reduction. *ALT*, 2013b.

Guermeur, Y. VC theory of large margin multi-category classifiers. *JMLR*, 8: 2551-2594, 2007.

Guermeur, Y. Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs. *Comm. Statist. Theory Methods*, 39(3):543-557, 2010.

Har-Peled, S. and Mendel, M. Fast construction of nets in low-dimensional metrics and their applications. *SIAM J. Comp.*, 35(5):1148-1184, 2006.

Kolmogorov, A. and Tikhomirov, V. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3-86, 1959.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist.*, 30(1):1-50, 2002.

Krauthgamer, R. and Lee, J. Navigating nets: Simple algorithms for proximity search. *SODA*, 2004.

Langford, J. and Beygelzimer, A. Sensitive error correcting output codes. *COLT*, 2005.

McShane, E. J. Extension of range of functions. *Bull. Amer. Math. Soc.*, 40 (12):837-842, 1934.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations Of Machine Learning*. The MIT Press, 2012.

Naor, A. and Schechtman, G. Planar earthmover is not in l_1. *SICOMP*, 37: 804-826, June 2007.

Papadimitriou, C. and Steiglitz, K. *Combinatorial optimization : algorithms and complexity*. Prentice Hall, 1998.

Rifkin, R. and Klautau, A. In defense of one-vs-all classification. *JMLR*, 5: 101-141, 2004.

von Luxburg, U. and Bousquet, O. Distance-based classification with lipschitz functions. *JMLR*, 5:669-695, 2004.

Weston, J. and Watkins, C. Support vector machines for multi-class pattern recognition. *ESANN* 99, 61-72, 1999.

Whitney, H. Analytic extensions of differentiable functions defined in closed sets. *Trans. Amer. Math. Soc.*, 36(1):63-89, 1934.

Zhang, T. Covering number bounds of certain regularized linear function classes. *JMLR*, 2:527-550, 2002.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *JMLR*, 5:1225-1251, 2004.

A Bayes near-optimality proof

Proof of Theorem 1. Since $\boldsymbol{\eta}$ is L-Lipschitz, given $x, x' \in \mathcal{X}$ we have

$$\begin{aligned} P(Y \neq Y' \,|\, x,x') &= \sum_{j \in \mathcal{Y}} \boldsymbol{\eta}_j(x)(1-\boldsymbol{\eta}_j(x')) \\ &\leq \sum_j \boldsymbol{\eta}_j(x)\left(1-\boldsymbol{\eta}_j(x)+Ld(x,x')\right) \\ &= \sum_j \boldsymbol{\eta}_j(x)\left(1-\boldsymbol{\eta}_j(x)\right)+Ld(x,x'). \end{aligned} \tag{18}$$

By the definition of the nearest neighbor classifier g_{NN} in (6) we have $\mathbb{E}_S[P(g_{\text{NN}}(X) \neq Y)] = \mathbb{E}_S[P(Y_{\pi_1(X)} \neq Y)]$, where the expectation is over the sample S determining g_{NN}. By (18) this error is bounded above by

$$\mathbb{E}_{S,X}[\sum_j \boldsymbol{\eta}_j(X)(1-\boldsymbol{\eta}_j(X))] + L\mathbb{E}_{S,X}[d(X, X_{\pi_1(X)})],$$

where now the expectation is over S and X. Denoting $k' = \operatorname{argmax}_j \boldsymbol{\eta}_j(X)$ and splitting the sum , the first term (which does not depend on S) satisfies

$$\begin{aligned} &\mathbb{E}_X[\boldsymbol{\eta}_{k'}(X)(1-\boldsymbol{\eta}_{k'}(X))] + \mathbb{E}_X[\sum_{j \neq k'} \boldsymbol{\eta}_j(X)(1-\boldsymbol{\eta}_j(X))] \\ &\leq \mathbb{E}_X[1-\boldsymbol{\eta}_{k'}(X)] + \mathbb{E}_X[\sum_{j \neq k'} \boldsymbol{\eta}_j(X)] \\ &= 2\mathbb{E}_X[1-\boldsymbol{\eta}_{k'}(X)] = 2P(g^*(X) \neq Y). \end{aligned}$$

It remains to bound $\mathbb{E}_{S,X}[d(X, X_{\pi_1(X)})]$ and we proceed exactly as in Ben-David & Shalev-Shwartz (2014). Let $\{C_1, \ldots, C_N\}$ be an ε-cover of $\mathcal{X}$ of cardinality $N = \mathcal{N}(\varepsilon, \mathcal{X}, d)$. Given a sample S, for $x \in C_i$ such that $S \cap C_i \neq \emptyset$ we have $d(x, X_{\pi_1(x)}) < \varepsilon$,

while for $x \in C_i$ such that $S \cap C_i = \emptyset$ we have $d(x, X_{\pi_1(x)}) \leq \operatorname{diam}(\mathcal{X}) = 1$, thus $\mathbb{E}_{S,X}[d(X, X_{\pi_1(X)})]$ is bounded above by

$$\begin{aligned} &\leq \mathbb{E}_S \left[\sum_{i=1}^{N} P(C_i) \left(\varepsilon \, \mathbb{1}_{\{S \cap C_i \neq \emptyset\}} + \mathbb{1}_{\{S \cap C_i = \emptyset\}} \right) \right] \\ &= \sum_{i=1}^{N} P(C_i) \left(\varepsilon \, \mathbb{E}_S \left[\mathbb{1}_{\{S \cap C_i \neq \emptyset\}} \right] + \mathbb{E}_S \left[\mathbb{1}_{\{S \cap C_i = \emptyset\}} \right] \right). \end{aligned}$$

Since $P(C_i)\mathbb{E}_S[\mathbb{1}_{S \cap C_i = \emptyset}] = P(C_i)(1 - P(C_i))^n \leq 1/en$ and $N = \mathcal{N}(\varepsilon, \mathcal{X}, d)$ we get

$$\begin{aligned} \mathbb{E}_{S,X}[d(X, X_{\pi_1(X)})] &\leq \varepsilon + \frac{\mathcal{N}(\varepsilon, \mathcal{X}, d)}{en} \\ &\leq \varepsilon + \frac{1}{en} \left(\frac{2}{\varepsilon} \right)^D. \end{aligned}$$

Setting $\varepsilon = 2n^{-\frac{1}{D+1}}$ concludes the proof. □

B Rademacher analysis proofs

Proof of inequality (10). Dudley's chaining integral (Dudley, 1967) bounds from above the Rademacher complexity $\mathcal{R}_n(\mathcal{H}_L)$ by

$$\inf_{\alpha > 0} \left(4\alpha + 12 \int_{\alpha}^{\infty} \sqrt{\frac{\log \mathcal{N}(t, \mathcal{H}_L, \|\cdot\|_\infty)}{n}} dt \right).$$

By Lemma 2 the integral can be bounded as follows:

$$\begin{aligned} &\int_{\alpha}^{\infty} \sqrt{\frac{\log \mathcal{N}(t, \mathcal{H}_L, \|\cdot\|_\infty)}{n}} dt \\ \leq \; &\int_{\alpha}^{\infty} \sqrt{\frac{1}{n} \left(\frac{16L}{t} \right)^D \log \left(\frac{5k}{t} \right)} dt \\ \leq \; &\int_{\alpha}^{\infty} \sqrt{\frac{\log 5k}{n} \left(\frac{16L}{t} \right)^D \left(\frac{1}{t} \right)} dt \\ = \; &\sqrt{\frac{\log 5k}{n}} (16L)^{D/2} \int_{\alpha}^{\infty} \left(\frac{1}{t} \right)^{(D+1)/2} dt \\ = \; &\sqrt{\frac{\log 5k}{n}} (16L)^{D/2} \left(\frac{2}{D-1} \right) \left(\frac{1}{\alpha^{(D-1)/2}} \right), \end{aligned}$$

where in the second inequality we used the fact that for $x \in (0, 1]$ and $c \geq e$ we have $\log(\frac{c}{x}) \leq \frac{\log c}{x}$. Choosing

$$\alpha^* = \left(9(16L)^D \frac{\log 5k}{n} \right)^{1/(D+1)}$$

yields the bound. □

Proof of Theorem 4. An adaptation[3] of Mohri et al. (2012, Theorem 4.5) to $\mathcal{H}_L$ states that with probability $1-\delta$, for all $L>0$, $h \in \mathcal{H}_L$,

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{\text{margin}}(h)] &\leq \widehat{\mathbb{E}}[\mathcal{L}_{\text{margin}}(h)] + 4\mathcal{R}_n(\mathcal{H}_L) \\ &+ \sqrt{\left(\frac{\log\log_2 2L}{n}\right)_+} + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.\end{aligned}$$

Since $\mathbb{1}_{\{u<0\}} \leq \mathcal{L}_{\text{margin}}(u)$ we have $P(g_h(X) \neq Y) \leq \mathbb{E}[\mathcal{L}_{\text{margin}}(h)]$. Since $\mathcal{L}_{\text{margin}}(u) \leq \mathcal{L}_{\text{cutoff}}(u)$ we can replace $\mathcal{L}_{\text{margin}}$ in the empirical loss by the loss function $\mathcal{L}_{\text{cutoff}}$. Bounding $\mathcal{R}_n(\mathcal{H}_L)$ using (10) concludes the proof. □

C Scale sensitive analysis proof

Proof of Theorem 5. An application[4] of Guermeur (2010, Theorem 1) states that with probability $1-\delta$, for all $L>0$, $h \in \mathcal{H}_L$,

$$\begin{aligned}P(g_h(X) \neq Y) &\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{h(X_i,Y_i)<1\}} \\ &+ \sqrt{\frac{2}{n}\left(2\log\mathcal{N}(1/4, \mathcal{H}_L, \|\cdot\|_\infty) + \ln\left(\frac{2L}{\delta}\right)\right)} + \frac{1}{n}.\end{aligned}$$

Applying the metric entropy bound in Lemma 2 proves the Theorem. □

D Approximate NN proofs

First, we will show that $\tilde{h}$ is indeed a 2η additive perturbation of h, i.e.

$$\|h - \tilde{h}\|_\infty \leq 2\eta. \tag{19}$$

Instead of working directly with (16) we consider the following L-Lipschitz extension

$$\begin{aligned}h(x,y) &= \frac{1}{2}\mathrm{T}_{[-1,1]}\left(\min_{S_1}\{\xi(Y_i,y) + Ld(X_i,x)\}\right) \\ &+ \frac{1}{2}\mathrm{T}_{[-1,1]}\left(\max_{S_1}\{\xi(Y_i,y) - Ld(X_i,x)\}\right),\end{aligned}$$

easily seen to induce the same classifier g_h as (16). Consider the first term (the second term is treated similarly) and its approximate version:

$$\tilde{h}(x,y) = \mathrm{T}_{[-1,1]}\left(\min_{S_1}\left\{\xi(Y_i,y) + L\tilde{d}(X_i,x)\right\}\right),$$

[3] essentially setting $\alpha = 1$ in Mohri et al. (2012) and doing the stratification on L instead
[4] setting $\gamma = 1$ in Guermeur (2010, Theorem 1) and doing the stratification on L instead

where $d \leq \tilde{d} \leq (1+\eta)d$, given in (15), is the approximate "'distance"' as provided by the approximate nearest neighbor. For notational convenience, denote

$$\begin{aligned} h(x,y) &= \mathrm{T}_{[\text{-}1,1]}(\min_i q_i(x,y)) \\ \tilde{h}(x,y) &= \mathrm{T}_{[\text{-}1,1]}(\min_i \tilde{q}_i(x,y)) \\ q_i(x,y) &= h_i(y) + r_i(x) \\ \tilde{q}_i(x,y) &= \tilde{h}_i(y) + \tilde{r}_i(x), \end{aligned}$$

where $h_i(y) = \xi(Y_i, y)$, $r_i(x) = Ld(X_i, x)$, and $\tilde{h}_i$, $\tilde{r}_i$ defined analogously.

Observe that if $\tilde{r}_i(x) > 2$ then $r_i(x) > 2/(1+\eta) \geq 2(1-\eta)$. In this case, since h has range in $[-1,1]$, the eventual application of truncation operator $\mathrm{T}_{[\text{-}1,1]}$ will force $\tilde{h}(x,y) - h(x,y) \leq 2\eta$. Hence, we may assume that $\tilde{r}_i(x) \leq 2$ and so $r_i(x) \leq 2$. It is straightforward to verify that for $a, b \in \mathbb{R}^n$ with $\max_{i\in[n]} |a_i - b_i| \leq \eta$, we have

$$\left| \mathrm{T}_{[\text{-}1,1]}(\min_i a_i) - \mathrm{T}_{[\text{-}1,1]}(\min_i b_i) \right| \leq \eta.$$

Thus, establishing $|q_i(x,y) - \tilde{q}_i(x,y)| \leq 2\eta$ for all $i \in [|S_1|]$ and $y \in \mathcal{Y}$ with $\tilde{r}_i(x), r_i(x) \leq 2$ suffices to prove the claim. Indeed, by (15) we have

$$|r_i(x) - \tilde{r}_i(x)| \leq |r_i(x) - (1+\eta)r_i(x)| \leq 2\eta.$$

Proof of Lemma 6. Suppose $\tilde{h} \in \mathcal{H}_{L,\eta}$. By the definition of $\mathcal{H}_{L,\eta}$, there exists an $h \in \mathcal{H}_L$ such that $\|\tilde{h} - h\|_\infty \leq \eta$. Let h' be some element in a minimal ε-cover of $\mathcal{H}_L$ so that $\|h - h'\|_\infty \leq \varepsilon$. Then

$$\|\tilde{h} - h'\|_\infty \leq \|\tilde{h} - h\|_\infty + \|h - h'\|_\infty \leq \varepsilon + \eta.$$

Hence,

$$\mathcal{N}(\varepsilon + \eta, \mathcal{H}_{L,\eta}, \|\cdot\|_\infty) \leq \mathcal{N}(\varepsilon, \mathcal{H}_L, \|\cdot\|_\infty),$$

whence the claim follows. □

E Dimensionality reduction proof

Proof of Theorem 7. Put $\tilde{S} = (\tilde{X}, Y)$. For $X_i \in X$ and $\tilde{X}_i \in \tilde{X}$, define $\delta_i(h) = h(X_i, Y_i) - h(\tilde{X}_i, Y_i)$. Then

$$\begin{aligned} \widehat{\mathcal{R}}_n(\mathcal{H}_L; S) &= \mathbb{E}\left[\sup_{h\in\mathcal{H}_L} \frac{1}{n} \sum_{i=1}^n \sigma_i h(X_i, Y_i) \,\middle|\, S \right] \\ &= \mathbb{E}\left[\sup_{h\in\mathcal{H}_L} \frac{1}{n} \sum_{i=1}^n \sigma_i \left(h(\tilde{X}_i, Y_i) - \delta_i(h) \right) \,\middle|\, S \right] \\ &\leq \widehat{\mathcal{R}}_n(\mathcal{H}_L; \tilde{S}) + \mathbb{E}\left[\sup_{h\in\mathcal{H}_L} \frac{1}{n} \sum_{i=1}^n \sigma_i \delta_i(h) \,\middle|\, S \right]. \end{aligned}$$

By (10), we have

$$\mathcal{R}_n(\mathcal{H}_L; \tilde{S}) \leq 2L \left(\frac{\log 5k}{n} \right)^{1/(\beta+1)}. \tag{20}$$

Since by construction h is L-Lipschitz in its first argument, we have

$$\left| \sum_{i=1}^{n} \sigma_i \delta_i(h) \right| \leq \sum_{i=1}^{n} |\delta_i(h)| \leq L \sum_{i=1}^{n} d(X_i, \tilde{X}_i) \leq L\alpha. \tag{21}$$

Our claimed bound follows from (20) and (21). □