A Bayes consistent 1-NN classifier

Aryeh Kontorovich and Roi Weiss
Computer Science Department
Ben Gurion University
Beer Sheva, Israel
{karyeh, roiwei}@cs.bgu.ac.il

July 2, 2014

Abstract

We show that a simple modification of the 1-nearest neighbor classifier yields a
strongly Bayes consistent learner. Prior to this work, the only strongly Bayes con-
sistent proximity-based method was the $k$-nearest neighbor classifier, for $k$ grow-
ing appropriately with sample size. We will argue that a margin-regularized 1-NN
enjoys considerable statistical and algorithmic advantages over the $k$-NN classi-
 fier. These include user-friendly finite-sample error bounds, as well as time- and
memory-efficient learning and test-point evaluation algorithms with a principled
speed-accuracy tradeoff. Encouraging empirical results are reported.

1 Introduction

The nearest neighbor (NN) classifier, introduced by Fix and Hodges in 1951, continues
to be a popular learning algorithm among practitioners. Despite the numerous sophis-
ticated techniques developed in recent years, this deceptively simple method continues
to “yield[] competitive results” [25] and inspire papers in “defense of nearest-neighbor
based […] classification” [3].

In the sixty years since the introduction of the nearest neighbor paradigm, a large
amount of theory has been developed for analyzing this surprisingly effective classification method. The first such analysis is due to [4], who showed that as the sample size
grows, the 1-NN classifier almost surely approaches an error rate $R \in [R^*, 2R^*(1 – R^*)]$, where $R^*$ is the Bayes-optimal rate. Although the 1-NN classifier is not in general Bayes consistent, taking a majority vote among the $k$ nearest neighbors does guarantee strong Bayes consistency, provided that $k$ increases appropriately in sample size [23, 5, 27].

The $k$-NN classifier in some sense addresses the Bayes consistency problem, but
presents issues of its own. A naive implementation involves storing the entire sample,
over which a linear-time search is performed when answering queries on test points.
For large samples sizes, this approach is prohibitively expensive in terms of storage
memory and computational runtime. To mitigate the memory concern, various condensing heuristics have been proposed [11, 7, 19, 26, 8]—of which only the one in [8] comes with any rigorous compression guarantees, and only for \(k = 1\); moreover, it is shown therein that the problem is ill-posed for \(k > 1\). Query evaluation on test points may be significantly sped up via an approximate nearest neighbor search [13, 2, 1, 9].

The price one pays for the fast approximate search is a degraded classification accuracy, and of the works cited, only [9] quantifies this tradeoff—and again, only for \(1\)-NN.

On the statistical front, one desires a classifier that provides an easily computable usable finite-sample generalization bound—one that the learner can evaluate based only on the observed sample so as to obtain a high-confidence error estimate. As we argue below, existing \(k\)-NN bounds fall short of this desideratum, and the few known usable bounds given in [24, 9, 8] are all for \(1\)-NN.

Motivated by the computational and statistical advantages that \(1\)-NN seems to enjoy over \(k\)-NN, this paper presents a strongly Bayes consistent \(1\)-NN classifier.

**Main results.** Our results build on the work of [9] and, more recently, [8]. Suppose we are given an iid training sample \(S\) consisting of \(n\) labeled points \((X_i, Y_i)\), with \(X_i\) residing in some metric space \(X\) and \(Y_i \in \{-1, 1\}\). For \(\varepsilon, \gamma > 0\), let us say that \(S\) is \((\varepsilon, \gamma)\)-separable if there is a sub-sample \(\tilde{S} \subset S\) such that (i) the \(1\)-NN classifier induced by \(\tilde{S}\) mislabels at most \(\varepsilon n\) points in \(S\) and (ii) every pair of opposite-labeled points in \(\tilde{S}\) is at least \(\gamma\) apart in distance. Obviously, a given sample \(S\) cannot be \((\varepsilon, \gamma)\)-separable for \(\varepsilon\) arbitrarily small and \(\gamma\) arbitrarily large. Every \(\gamma > 0\) determines some minimum feasible \(\varepsilon^* = \varepsilon^*(\gamma)\) and a corresponding \(\varepsilon^*\)-consistent, \(\gamma\)-separable sub-sample \(S^*(\gamma) \subset S\).

Margin-based generalization bounds were presented in [9, 8], with \(\varepsilon\) corresponding to empirical error and \(\gamma\) to the margin. Schematically, these bounds are of the form

\[
\text{gen-err}_n(\varepsilon, \gamma) \leq \text{empirical}_n(\varepsilon, \gamma) + \text{complexity}_n(\gamma),
\]

where gen-err is the generalization error of the \(1\)-NN classifier induced by an \(\varepsilon\)-consistent, \(\gamma\)-separable \(\tilde{S} \subset S\), and the two terms on the right-hand side correspond roughly to sample error and hypothesis complexity. The approach proposed in [9, 8] suggests computing \(\varepsilon^*(\gamma)\) for each \(\gamma > 0\) and minimizing the right-hand side of (1) over \(\gamma\) to obtain \(\gamma^*_n\). Indeed, the chief technical contribution of those works consists providing efficient algorithms for computing \(\varepsilon^*(\gamma)\), \(S^*(\gamma)\), and \(\gamma^*_n\). In contrast, the present paper deals with the statistical aspects of this procedure. Our main contribution is Theorem 2 which shows that the \(1\)-NN classifier induced by \(S^*(\gamma^*_n)\) is strongly Bayes consistent. Denoting this classifier by \(h_n\), our main result is formally stated as follows:

\[
P(h_n(X) \neq Y \mid (X_1, Y_1), \ldots, (X_n, Y_n)) \xrightarrow{a.s.} R^*,
\]

where

\[
R^* = \inf_{h : X \to \{-1, 1\}} \mathbb{P}(h(X) \neq Y)
\]
is the Bayes-optimal error. This is the first consistency result (strong or otherwise) for an algorithmically efficient 1-NN classifier.

**Related work.** Following the pioneering work of Cover and Hart [4], it was shown by [5][27] that the $k$-NN classifier is strongly Bayes consistent. A representative result for the Euclidean space $\mathcal{X} = \mathbb{R}^d$ states that if $k \to \infty$ and $k/n \to 0$, then for all $\varepsilon > 0$ and $n > n_0(\varepsilon, k)$,

$$
P(R(h_{k-NN}) > R^* + \varepsilon) \leq 2 \exp \left( -\frac{n\varepsilon^2}{5184\sigma_d^2} \right),$$  \hspace{1cm} (2)

where $\sigma_d < \left( 1 + 2/\sqrt{2 - \sqrt{3}} \right)^d$ is the minimum number of origin-centered cones of angle $\pi/6$ that cover $\mathbb{R}^d$ (this result, among many others, is proved in [6]). Given the inherently Euclidean nature of $\sigma_d$, (2) does not seem to readily extend to more general metric spaces. More recently, [21] (essentially) showed that

$$
\mathbb{E}[R(h_{k-NN})] \leq \left( 1 + \sqrt{8/k} \right) R^* + (6L + k) n^{-1/(d+1)}$$  \hspace{1cm} (3)

for metric spaces $\mathcal{X}$ with unit diameter and doubling dimension $d$ (defined below), where $L$ is the Lipschitz constant of $\eta : \mathcal{X} \to [0,1]$ defined by $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$.

Although (2,3) are both finite-sample bounds, they do not enable a practitioner to compute a numerical generalization error estimate for a given training sample. Both are stated in terms of the unknown Bayes-optimal rate $R^*$, and (3) additionally depends on $L$, a property of the unknown distribution. In particular, (2) and (3) do not allow for a data-dependent selection of $k$, which must be tuned via cross-validation. The asymptotic expansions in [22][18] likewise do not provide a computable finite-sample bound.

An entire chapter in [6] is devoted to condensed and edited NN rules. In the terminology of this paper, this amounts to extracting a sub-sample $\tilde{S}$ and predicting via the 1-NN classifier induced by that $\tilde{S}$. Assuming a certain sample compression rate and an oracle for choosing an optimal fixed-size $\tilde{S}$, this scheme is shown to be weakly Bayes consistent. The generalizing power of sample compression was independently discovered by [15], and later elaborated upon by [10]. In the context of NN classification, [6] lists various condensing heuristics (which have no performance guarantees) and also leaves open the algorithmic question how to minimize the empirical loss over all subsets of a given size.

The first substantial departure from the $k$-NN paradigm was proposed by [24], with the straightforward but far-reaching observation that the 1-NN classifier is, in some sense, equivalent to interpreting the labeled sample $\{(X_i, Y_i) : i \in [n]\}$ as $n$ evaluations of a real-valued target function $f$, computing its Lipschitz extension $f^*$ from the sample points to all of $\mathcal{X}$, and then classifying test points by $\text{sign}(f^*(\cdot))$. Following up, [9] obtained bounds on the fat-shattering dimension of Lipschitz functions in doubling spaces and gave the first risk decay as $\tilde{O}(1/\sqrt{n})$ as opposed to $1/n^d$. More recently, the existence of a margin was leveraged to give nearly optimal sample compression bounds [8].
2 Preliminaries

Metric Spaces. Throughout this paper, our instance space $\mathcal{X}$ will be endowed with a bounded metric $\rho$, which we will normalize to have unit diameter $\text{diam}(\mathcal{X}) := \sup_{x,x' \in \mathcal{X}} \rho(x, x') = 1$.

A function $f : \mathcal{X} \rightarrow \mathbb{R}$ is said to be $L$-Lipschitz if $|f(x) - f(x')| \leq L\rho(x, x')$ for all $x, x' \in \mathcal{X}$. The Lipschitz constant of $f$, denoted $\|f\|_{L^*}$, is the smallest $L$ for which $f$ is $L$-Lipschitz. The collection of all $L$-Lipschitz $f : \mathcal{X} \rightarrow [-1, 1]$ will be denoted by $\mathcal{F}_L$. The distance between two sets $A, B \subseteq \mathcal{X}$ is defined by $\rho(A, B) = \inf_{x \in A, x' \in B} \rho(x, x')$. For a metric space $(\mathcal{X}, \rho)$, let $\lambda$ be the smallest value such that every ball in $\mathcal{X}$ can be covered by $\lambda$ balls of half the radius. The doubling dimension of $\mathcal{X}$ is $\text{ddim}(\mathcal{X}) := \log_2 \lambda$. A metric is doubling when its doubling dimension is finite. We will denote $d := \text{ddim}(\mathcal{X}) < \infty$.

Learning model. We work in the standard agnostic learning model \cite{16, 21}, whereby the learner receives a sample $S$ consisting of $n$ labeled examples $(X_i, Y_i)$, drawn iid from an unknown distribution over $\mathcal{X} \times \{-1, 1\}$. All subsequent probabilities and expectations will be with respect to this distribution. Based on the training sample $S$, the learner produces a hypothesis $h : \mathcal{X} \rightarrow \{-1, 1\}$, whose empirical error is defined by $\hat{R}_n(h) = n^{-1} \sum_{i=1}^n 1_{\{h(X_i) \neq Y_i\}}$ and whose generalization error is defined by $R(h) = \mathbb{P}(h(X) \neq Y)$. The Bayes-optimal classifier, $h^*$, is defined by $h^*(x) = \arg\max_{y \in \{-1, 1\}} \mathbb{P}(Y = y | X = x)$ and $R^* := R(h^*) = \inf \{R(h)\}$, where the infimum is over all measurable hypotheses. A learning algorithm mapping a sample $S$ of size $n$ to a hypothesis $h_n$ is said to be strongly Bayes consistent if $R(h_n) \xrightarrow{n \rightarrow \infty} R^*$ almost surely.

Sub-sample, margin, and induced 1-NN. In a slight abuse of notation, we will blur the distinction between $S \subseteq \mathcal{X}$ as a collection of points in a metric space and $S \in (\mathcal{X} \times \{-1, 1\})^n$ as a sequence of labeled examples. Thus, the notion of a sub-sample $\tilde{S} \subseteq S$ partitioned into its positively and negatively labeled subsets as $\tilde{S} = \tilde{S}_+ \cup \tilde{S}_-$ is well-defined. The margin of $\tilde{S}$, defined by $\text{marg}(\tilde{S}) = \rho(\tilde{S}_+, \tilde{S}_-)$, is the minimum distance between a pair of opposite-labeled points. A sub-sample $\tilde{S}$ naturally induces the 1-NN classifier $\tilde{h}_{\tilde{S}}$, via $\tilde{h}_{\tilde{S}}(x) = \text{sign}(\rho(x, \tilde{S}_-) - \rho(x, \tilde{S}_+))$.

Margin risk. For a given sample $S$ of size $n$ any $\gamma > 0$ and measurable $f : \mathcal{X} \rightarrow \mathbb{R}$, we define the margin risk $R_{\gamma}(f) = \mathbb{P}(f(X) < \gamma)$ and its empirical version $\hat{R}_{n, \gamma}(f) = n^{-1} \sum_{i=1}^n 1_{\{Y_i f(X_i) < \gamma\}}$. When $\gamma = 0$, we omit it from the subscript; thus, e.g., $R(f) = \mathbb{P}(f(X) < 0)$, which agrees with the definitions $R(h)$ and $\hat{R}_n(h)$ above for binary-valued $h$.

\footnote{This assumption is not really restrictive, as any finite sample will be contained in some ball. The situation is analogous to margin-based analysis of Euclidean hyperplanes, where the quantity of interest is the ratio between data diameter and separation margin.}
3 Learning Algorithm: Regularized 1-NN

This section is mainly provided to cast known results (or their minor modifications) in the terminology of this paper. As the main contribution of this paper is a Bayes-consistency analysis of a particular learning algorithm, we must first provide the details of the latter. The learning algorithm in question is essentially the one given in [9]. Our point of departure is the connection made by [24] between Lipschitz functions and 1-NN classifiers.

Theorem 1 ([24]). If \( \tilde{S} \) is a sub-sample with \( \text{marg}(\tilde{S}) \geq \gamma \), then there is an \( f \in F_2 \) such that

\[
h_{\tilde{S}}(x) = \text{sign}(f(x)) \quad \text{for all } x \in X.
\]

More explicitly, \( f \) is a Lipschitz extension of \( \tilde{S} \), satisfying

\[
f(x) = f_{\tilde{S}}(x) = \begin{cases} +\gamma, & \text{if } x \in \tilde{S}_+ \\ -\gamma, & \text{if } x \in \tilde{S}_- \end{cases}.
\]  

We will only consider members of \( F_2 \) that are Lipschitz-extensions of \( \gamma \)-separable sub-samples and will never need to actually calculate these explicitly; their only purpose is to facilitate the analysis. In line with the Structural Risk Minimization (SRM) paradigm, our learning algorithm consists of minimizing the penalized margin risk,

\[
\hat{R}^{\text{PEN}}_{n,\gamma}(f) = \hat{R}_{n,\gamma}(f) + r^{\text{PEN}}(n, \gamma),
\]

where

\[
r^{\text{PEN}}(n, \gamma) = \frac{2 \log^+ \frac{1}{\gamma} \left( \frac{c_4 d}{n} \right)}{\gamma} + \sqrt{\frac{2c_1}{d+1} \log (n/c_d) + 4c_1 \log \log \frac{1}{\gamma}}
\]

and \( c_1, c_d \) are explicitly computable constants, the latter depending only on \( d \). The form of the penalty term will be motivated by the analysis in the sequel.

This optimization is performed via two nested routines: the inner one minimizes \( \hat{R}^{\text{PEN}}_{n,\gamma}(f) \) over \( f \in F_2 \) for a fixed \( \gamma \), while the outer one minimizes over \( \gamma > 0 \). Since this is a very slight modification of the SRM procedure proposed and analyzed in [9], we will give a high-level sketch.

Inner routine: optimizing over \( f \in F_2 \). By Theorem 1 minimizing \( \hat{R}^{\text{PEN}}_{n,\gamma}(f) \) over \( f \in F_2 \) for a fixed \( \gamma \) is equivalent to seeking a \( \gamma \)-separable \( S \subset S \) whose induced 1-NN classifier \( h_{\tilde{S}} \) makes the fewest mistakes on \( S \) (see Algorithm 1). The algorithm invokes a minimum vertex cover routine, which by König’s theorem is equivalent to maximum matching for bipartite graphs, and is computable in randomized time \( O(n^{2.376}) \) [17].

Outer loop: minimizing over \( \gamma > 0 \). Although \( \gamma \) takes on a continuum of values, we need only consider those induced by distances between opposite-labeled points in \( S \), of which there are \( O(n^2) \). For each candidate \( \gamma \), Algorithm 1 computes the optimal \( f_{\gamma} \in F_2 \). Let \( \gamma^*_n \) be a minimizer of \( \hat{R}^{\text{PEN}}_{n,\gamma}(f_{\gamma}) \), with corresponding \( f^{\text{PEN}}_{\gamma^*_n} \equiv f_{\gamma^*_n}^*:

\[
\hat{R}^{\text{PEN}}_{n,\gamma^*_n} := \inf_{\gamma > 0} \inf_{f \in F_2} \hat{R}^{\text{PEN}}_{n,\gamma}(f) = \hat{R}^{\text{PEN}}_{n,\gamma^*_n}(f^{\text{PEN}}_{\gamma^*_n}).
\]
Algorithm 1 minimizing \( \hat{R}_{\text{PEN}}(f) \) over \( f \in \mathcal{F}_2 \) for a fixed \( \gamma \)

1: function INNER\((S, \gamma)\)
2: construct bipartite graph \( G = (S_+, S_-, E) \) with
   \[
   E = \{(x, x') : x \in S_+, x' \in S_-, \rho(x, x') < \gamma\}
   \]
3: compute minimum vertex cover \( C = C_+ \cup C_- \) for \( G \)
4: return \( \tilde{S} = S \setminus C \)
5: end function

The total runtime for computing \( \gamma_n^* \) and \( f_{\text{PEN}}^n \) is \( O(n^{4.376}) \), which may be considerably sped up if we are willing to tolerate a small approximation factor [9].

4 Consistency proof

In this section we prove the main technical result of this paper:

Theorem 2. With probability one over the random sample \( S \) of size \( n \),
\[
\lim_{n \to \infty} R(f_{\text{PEN}}^n) = R^*.
\]

As the proof is somewhat involved, we will break it up into high-level steps. The basic idea is to decompose the excess risk into two terms,
\[
R(f_{\text{PEN}}^n) - R^* = \left( R(f_{\text{PEN}}^n) - \hat{R}_{\text{PEN}}^n \right) + \left( \hat{R}_{\text{PEN}}^n - R^* \right) = (I) + (II),
\]
and to show that both decay to 0 almost surely. All omitted proofs are given in the Appendix.

4.1 The term (I)

Intuitively, for each \( \gamma, L > 0 \), we would like to prove a deviation estimate on \( |R_{\gamma}(f) - \hat{R}_{n,\gamma}(f)| \) uniformly over all \( f \in \mathcal{F}_L \). We find it most convenient to proceed using Rademacher complexities but these require a loss that is Lipschitz-continuous in \( \gamma \) — and \( \hat{R}_{n,\gamma}(f) \) is not even continuous (it is lower-semicontinuous in \( \gamma \) for a fixed \( f \)). We overcome this technical hurdle by introducing a surrogate loss \( \Phi_\xi \) and surrogate risk \( L_\xi \) as follows.

Surrogate loss. For \( 0 < \xi \leq 1 \) define the surrogate loss function \( \Phi_\xi(u) : \mathbb{R} \to [0, 1] \),
\[
\Phi_\xi(u) = \begin{cases} 1 & \text{if } u \leq 1 - \xi, \\ 0 & \text{if } u \geq 1, \\ (1 - u)/\xi & \text{otherwise}, \end{cases}
\]

\[2\] An alternative, though somewhat messier route, would be to use fat-shattering dimension, as in [9].
and its associated empirical and expected surrogate risks, 

\[ \hat{L}_{\xi}(f) = \frac{1}{n} \sum_{i=1}^{n} \Phi_{\xi}(Y_i f(X_i)), \quad L_{\xi}(f) = \mathbb{E}[\Phi_{\xi}(Y f(X))]. \]  

(10)

At this point, it appears as though we have three free parameters: \( \gamma, L, \) and \( \xi \). However, we will tie them together via the relation \( L = \frac{2}{\gamma} \) and a common (double) stratification scheme. For \( n, l \in \mathbb{N} \) put

\[ \xi_{n,l} = \frac{1}{l \cdot n_d}, \]  

(11)

\[ \gamma_{n,l+1} = \gamma_{n,l} (1 - \xi_{n,l}), \quad \gamma_{n,1} = 1, \]  

(12)

\[ \varepsilon_{n,l} = \frac{2}{\gamma_{n,l} \xi_{n,l} n_d^2} + \sqrt{4c_1 \log \left( \frac{1}{\gamma_{n,l}} \right)}, \]  

(13)

where

\[ n_d = \left( \frac{n}{c_d} \right)^{\frac{1}{d+1}}. \]  

(14)

This enables us to obtain a uniform deviation estimate:

**Lemma 3.** For all \( n \in \mathbb{N} \) and \( \varepsilon > 0 \),

\[ \mathbb{P}\left( \exists l \in \mathbb{N} : \sup_{f \in F_{L_n,l}} \left| L_{\xi_{n,l}}(f) - \hat{L}_{\xi_{n,l}}(f) \right| > \varepsilon + \varepsilon_{n,l} \right) \leq \frac{n^2}{6} \exp \left( -\frac{n \varepsilon^2}{c_1} \right). \]

We proceed with showing the term (I) decays to zero almost surely. By Theorem 1 we may fix \( f_{\text{Pen}}^\text{rsn} \) to belong to \( F_2 \). With \( \gamma_n^* = \text{margin}(\tilde{S}) \) let \( l_n^-, l_n^+ \in \mathbb{N} \) to be the consecutive margin indexes in the stratification grid (12) such that

\[ \gamma_n^* \in [\gamma_{n,l_n^-}, \gamma_{n,l_n^+}), \quad l_n^- - l_n^+ = 1, \quad \forall n \in \mathbb{N} \]

and abbreviate \( \gamma_n^+ = \gamma_{n,l_n^+}, \gamma_n^- = \gamma_{n,l_n^-} \). Setting

\[ g_n^\text{rsn} = (1/\gamma_n^*) f_{\text{Pen}}^\text{rsn} \in F_2/\gamma_n^* \subset F_2/\gamma_n^- \]

Note that we have

\[ \gamma_{n,l} = \frac{1}{n_d} \sum_{j=1}^{l-1} \left( 1 - \frac{1}{j + n_d} \right) \approx \exp \left( -\frac{1}{n_d} \sum_{j=1}^{l-1} \frac{1}{j+n_d} \right) \approx \exp \left( -\frac{(l-1) - \frac{1}{2}}{n_d(1 - \frac{1}{n_d})} \right) \]

and thus

\[ \xi_{n,l} \approx \frac{1}{n_d \log \left( \frac{1}{\gamma_{n,l}} \right)} \]

which implies the form of \( f_{\text{Pen}}^\text{rsn}(n, \gamma) \) in (6).

\footnote{Note that we have \( \gamma_{n,l} \approx 1/\gamma_{n,l} \approx n_d \log \left( \frac{1}{\gamma_{n,l}} \right) \) which implies the form of \( f_{\text{Pen}}^\text{rsn}(n, \gamma) \) in (6).}

\footnote{in this section we use the symbol \( f \) for functions in \( F_2 \) and \( g \) for functions in \( F_L \) with arbitrary \( L \).}
yields

\[(*) = \mathbb{P} \left( R(f_n^{\text{PEN}}) - \hat{R}_{n}^{\text{PEN}} > \varepsilon \right) \]  
\[= \mathbb{P} \left( R(f_n^{\text{PEN}}) - \hat{R}_{n,\gamma_n^*}^{\text{PEN}} - r_{\text{ens}}(n, \gamma_n^*) > \varepsilon \right) \]  
\[= \mathbb{P} \left( R(g_n^{\text{PEN}}) - \hat{R}_{n,1}^{\text{PEN}} > \varepsilon + r_{\text{ens}}(n, \gamma_n^*) \right). \]

Since \( \xi_{n,l} \to 0 \) monotonically, we have

\[R(g_n^{\text{PEN}}) \leq \mathcal{L}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) \]
\[\hat{R}_{n,1}(g_n^{\text{PEN}}) \geq \hat{\mathcal{L}}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) \]
\[r_{\text{ens}}(n, \gamma_n^*) \geq r_{\text{ens}}(n, \gamma_n^+). \]

Thus

\[(*) \leq \mathbb{P} \left( \mathcal{L}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) - \hat{\mathcal{L}}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) > \varepsilon + r_{\text{ens}}(n, \gamma_n^+) \right). \]

Note that

\[\left| \mathcal{L}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) - \mathcal{L}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) \right| \leq \sup_{u \in [-1,1]} \left| \Phi_{\xi_{n,l}^+}^{\text{PEN}}(u) - \Phi_{\xi_{n,l}^+}^{\text{PEN}}(u) \right| \]
\[= \frac{\xi_{n,l}^+ - \xi_{n,l}^-}{\xi_{n,l}^+} \to 0, \]
where the limit holds by the definition of \( \xi_{n,l} \). Taking \( n \) sufficiently large, we have

\[(*) \leq \mathbb{P} \left( \mathcal{L}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) - \hat{\mathcal{L}}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) \right) > \varepsilon + r_{\text{ens}}(n, \gamma_n^+) \]
\[\leq \mathbb{P} \left( \mathcal{L}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) - \hat{\mathcal{L}}_{\xi_{n,l}^+}^{\text{PEN}}(g_n^{\text{PEN}}) \right) > \varepsilon/2 + r_{\text{ens}}(n, \gamma_n^+) \]
\[\leq \mathbb{P} \left( \sup_{g \in \mathcal{F}_{2/\gamma_n^+}} \left| \mathcal{L}_{\xi_{n,l}^+}^{\text{PEN}}(g) - \hat{\mathcal{L}}_{\xi_{n,l}^+}^{\text{PEN}}(g) \right| > \varepsilon/2 + r_{\text{ens}}(n, \gamma_n^+) \right) \]
\[\leq \mathbb{P} \left( \exists l \in \mathbb{N} : \sup_{g \in \mathcal{F}_{2/\gamma_n^+}} \left| \mathcal{L}_{\xi_{n,l}^+}(g) - \hat{\mathcal{L}}_{\xi_{n,l}^+}(g) \right| > \varepsilon/2 + r_{\text{ens}}(n, \gamma_n^+) \right) \]
\[\leq \frac{\pi^2}{6} \exp \left( -\frac{n \varepsilon^2}{4c_1} \right), \]

where the last inequality is by Lemma 3.

### 4.2 The term (II)

We will need to approximate the Bayes optimal risk by margin risks:

**Lemma 4.** For every \( \varepsilon > 0 \) there is a \( \gamma > 0 \) such that

\[\inf_{f \in \mathcal{F}_2} R_{\gamma}(f) - R^* < \varepsilon.\]
In particular,

\[ R^* = \lim_{\gamma \to 0} \inf_{f \in F_2} R_\gamma(f). \]  

Since (18) holds for any sequence \( \gamma_n \to 0 \), we will assume without loss of generality that it is a subsequence of the stratification grid (12). Hence, for all \( \varepsilon > 0 \), there is a \( \bar{\gamma}^+ \) with a corresponding \( \tilde{\gamma}^+ \) such that

\[ R_{\bar{\gamma}^+}(\tilde{\gamma}^+) \leq \inf_{f \in F_2} R_{\bar{\gamma}^+}(f) + \varepsilon / 4. \]

Fix such a \( \bar{\gamma}^+ \) and let \( \tilde{\gamma}^- \) be the following margin in the grid (12) (so \( \bar{\gamma}^- < \bar{\gamma}^+ \) are consecutive) with corresponding indexes \( \tilde{\gamma}^+ \). Thus, it suffices to show that

\[ \limsup_{n \to \infty} \hat{R}_{n} \leq R_{\bar{\gamma}^+}(\tilde{\gamma}^+) \]

almost surely. Now Algorithm 1 provides an \( f_{n} \) such that

\[ R_{n,\tilde{\gamma}^+}(f_{n}) \leq \hat{R}_{n,\tilde{\gamma}^-}(\tilde{\gamma}^+). \]

Thus, similarly as for the first term, we have

\[ (**) = \mathbb{P}\left( \hat{R}_{n} - \inf_{f \in F_2} R_{\bar{\gamma}^+}(f) > \varepsilon \right) \]

\[ \leq \mathbb{P}\left( \hat{R}_{n} - R_{\bar{\gamma}^+}(\tilde{\gamma}^+) > 3\varepsilon / 4 \right) \]

\[ \leq \mathbb{P}\left( \hat{R}_{n,\tilde{\gamma}^-} - R_{\bar{\gamma}^+}(\tilde{\gamma}^+) > 3\varepsilon / 4 - r_{n,\tilde{\gamma}^-} \right) \]

Take \( n \) large enough so that \( r_{n,\tilde{\gamma}^-} \leq \varepsilon / 4 \) and rescale

\[ \tilde{\gamma}^+ = (1/\bar{\gamma}^+) \tilde{\gamma}^+ \in F_2/\bar{\gamma}^+ \]

to obtain

\[ (**) \leq \mathbb{P}\left( \hat{R}_{n,\tilde{\gamma}^-/\bar{\gamma}^+} - R_{1}(\tilde{\gamma}^+) > \varepsilon / 2 \right) = (**). \]

The next step is to note that, for \( \gamma \) fixed to be 1, the margin loss \( R_{\gamma=1}(\cdot) \) is well-approximated by surrogate losses:

**Lemma 5.** For every \( L > 0 \) and \( f \in F_L \)

\[ \lim_{n \to \infty} \sup_{t \in \mathbb{N}} \left| \mathcal{L}_{\xi_n,i}(f) - R_{1}(f) \right| = 0. \]  

By Lemma 5 we can take \( n \) large enough so that

\[ \left| \mathcal{L}_{\xi_n,i}(\tilde{\gamma}^+) - R_{1}(\tilde{\gamma}^+) \right| \leq \varepsilon / 4, \]
and since by definition $\forall l \in \mathbb{N}$

\[
\frac{\gamma_{n,l+1}}{\gamma_{n,l}} = 1 - \xi_{n,l}
\]

we have

\[
\hat{R}_{n,l}/\tilde{\gamma}^+(\hat{g}^+) \leq \hat{L}_{\xi_{n,l}}(\tilde{g}^+).
\]

Putting these together we finally get

\[
(\text{***}) \leq \mathbb{P}\left( \hat{L}_{n,\xi_{n,l}}(\tilde{g}^+) - \mathcal{L}_{n,\xi_{n,l}}(\tilde{g}^+) > \varepsilon/4 \right) \\
\leq \mathbb{P}\left( \sup_{g \in \mathcal{F}_{2/\tilde{\gamma}^+}} \left| \hat{L}_{n,\xi_{n,l}}(g) - \mathcal{L}_{n,\xi_{n,l}}(g) \right| > \varepsilon/4 \right) \leq ce^{-\frac{\varepsilon^2}{16}},
\]

analogously to the bound on term (I).

5 Experiments

We ran simulations with a twofold purpose: (a) to ascertain the convergence of various classifier risks to the Bayes optimal risk and to compare their rates of convergence and (b) to compare the actual runtimes of the various algorithms. To this end, we took $\mathcal{X} = \mathbb{R}^2$ endowed with the Euclidean metric $\rho = \| \cdot \|_2$, and defined the following joint distribution on $\mathcal{X} \times \{-1, 1\}$. A point $(x_1, x_2) \in \mathcal{X}$ is sampled by drawing $T \in [0, 2\pi]$ uniformly at random and defining

\[
x_1(T) = A\sqrt{T} \cos(\omega T), \\
x_2(T) = A\sqrt{T} \sin(\omega T)
\]

for some fixed $\omega$. The label $Y \in \{-1, 1\}$ is drawn according to

\[
\eta(T) = \mathbb{P}(Y = 1 \mid T) = \frac{1 + \cos(\omega T)}{2}.
\]

(See Figure 1 for an illustration.)

We compared four classifiers: $k^\ast$-NN (the $k$-NN classifier with $k$ optimized by cross-validation), SVM (support vector machine with the RBF kernel whose bandwidth and regularization penalty were optimized by cross-validation), CV-1-NN (margin-regularized 1-NN with $\gamma$ tuned by cross-validation), and SRM-1-NN (the 1-NN classifier described in Section 3 using a greedy vertex cover heuristic rather than the exact matching algorithm). Their runtime and generalization performance are summarized in Figure 2. Our proposed algorithm, SRM-1-NN, emerges competitive by both criteria.
Figure 1: The distribution with $A = 5$ and $\omega = 3$.

Figure 2: Comparing generalization error and running time vs. number of samples.

(a) Generalization error vs. number of samples. CV-1-NN is uniformly dominant, but for large sample sizes SRM-1-NN catches up. Unregularized 1-NN is included for reference; it is clearly not Bayes consistent.

(b) Running time vs. number of samples. SRM-1-NN enjoys a clear time advantage over the other methods involving cross-validation.

References

[1] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122, 2008.

[2] Alina Beygelzimer, Sham Kakade, and John Langford. Cover trees for nearest neighbor. In *ICML ’06: Proceedings of the 23rd international conference on Machine learning*, pages 97–104, New York, NY, USA, 2006. ACM.

[3] Oren Boiman, Eli Shechtman, and Michal Irani. In defense of nearest-neighbor based image classification. In *CVPR*, 2008.

[4] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.
[5] Luc Devroye and László Györfi. *Nonparametric density estimation: the L_1 view*. Wiley Series in Probability and Mathematical Statistics: Tracts on Probability and Statistics. John Wiley & Sons, Inc., New York, 1985.

[6] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of Applications of Mathematics (New York). Springer-Verlag, New York, 1996.

[7] W. Gates. The reduced nearest neighbor rule. *IEEE Transactions on Information Theory*, 18:431–433, 1972.

[8] Lee-Ad Gottlieb and Aryeh Kontorovich. Near-optimal sample compression for nearest neighbors, arxiv:1404.3368. 2014.

[9] Lee-Ad Gottlieb, Leonid Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. In COLT, pages 433–440, 2010.

[10] Thore Graepel, Ralf Herbrich, and John Shawe-Taylor. Pac-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

[11] Peter E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3):515–516, 1968.

[12] Aryeh Kontorovich and Roi Weiss. Maximum margin multiclass nearest neighbors. In ICML, 2014.

[13] Robert Krauthgamer and James R. Lee. Navigating nets: Simple algorithms for proximity search. In 15th Annual ACM-SIAM Symposium on Discrete Algorithms, pages 791–801, January 2004.

[14] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

[15] Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability, unpublished. 1986.

[16] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations Of Machine Learning*. The MIT Press, 2012.

[17] Marcin Mucha and Piotr Sankowski. Maximum matchings via gaussian elimination. In FOCS ’04: Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science, pages 248–255, Washington, DC, USA, 2004. IEEE Computer Society.

[18] Demetri Psaltis, Robert R. Snapp, and Santosh S. Venkatesh. On the finite sample performance of the nearest neighbor classifier. *IEEE Transactions on Information Theory*, 40(3):820–837, 1994.

[19] G. L. Ritter, H. B. Woodruff, S. R. Lowry, and T. L. Isenhour. An algorithm for a selective nearest neighbor decision rule. *IEEE Transactions on Information Theory*, 21:665–669, 1975.

[20] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill, 1987.

[21] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[22] Robert R. Snapp and Santosh S. Venkatesh. Asymptotic expansions of the k nearest neighbor risk. *Ann. Statist.*, 26(3):850–878, 1998.

[23] Charles J. Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4):595–620, 1977.

[24] Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, 2004.

[25] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[26] D. R. Wilson and T. R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine Learning*, 38:257–286, 2000.

[27] Lin Cheng Zhao. Exponential bounds of mean error for the nearest neighbor estimates of regression functions. *J. Multivariate Anal.*, 21(1):168–178, 1987.
A Appendix

A.1 Proof of Lemma 3

We first need the uniform convergence lemma:

**Lemma 6.** For $0 < \varepsilon, 0 < \xi < 1, 0 < L$ and iid samples $S$ of size $n$,

$$
\Pr \left( \sup_{f \in \mathcal{F}_L} |\mathcal{L}_\xi(f) - \hat{\mathcal{L}}_{n,\xi}(f) - \mathcal{L}_\xi(f) + \epsilon\right) \leq \exp \left( \frac{-n\varepsilon^2}{c_1}\right),
$$

where the Rademacher complexity $\mathcal{R}_n(\mathcal{L}_\gamma \circ \mathcal{F}_L)$ is bounded from above by:

$$
\mathcal{R}_{n,L,\xi} := c_d \frac{L}{\xi} \left( \frac{1}{n} \right)^{\frac{d+1}{2}}.
$$

**Proof of Lemma 6.** Equation (21) is restatement of \cite[Theorem 3.1]{16}. Note that $\Phi_\xi : \mathbb{R} \to [0, 1]$ is $\frac{1}{\xi}$-Lipschitz. Thus, by Talagrand’s contraction lemma \cite{14},

$$
\mathcal{R}_n(\mathcal{L}_\xi \circ \mathcal{F}_L) \leq \frac{L}{\xi} \mathcal{R}_n(\mathcal{F}_1).
$$

The Rademacher complexity for the class of 1-Lipschitz functions on a $\text{diam}(\mathcal{X}) = 1$ $d$-doubling metric space $\mathcal{R}_n(\mathcal{F}_1)$ is bounded as in \cite{12}.

**Proof.** Following proof idea in \cite[Theorem 18.2]{6}, a union bound yields

$$
(*) := \Pr \left( \exists l \in \mathbb{N} : \sup_{f \in \mathcal{F}_{n,l}} \left| \mathcal{L}_{\xi,n,l}(f) - \hat{\mathcal{L}}_{n,\xi,n,l}(f) \right| > \varepsilon + \varepsilon_{n,l} \right)
$$

$$
\leq \sum_{l=1}^{\infty} \Pr \left( \sup_{f \in \mathcal{F}_{n,l}} \left| \mathcal{L}_{\xi,n,l}(f) - \hat{\mathcal{L}}_{n,\xi,n,l}(f) \right| > \varepsilon + \varepsilon_{n,l} \right).
$$

Bounding each term in the sum by Lemma 6 we have,

$$
(*) \leq \sum_{l=1}^{n_1} e^{-\frac{n(\varepsilon + \varepsilon_{n,l-1})^2}{c_1}} \leq e^{-\frac{\varepsilon^2}{c_1}} \sum_{l=1}^{\infty} e^{-\frac{n(n-1)\varepsilon^2}{c_1}} = \frac{n^2}{6} e^{-\frac{\varepsilon^2}{c_1}}.
$$

**A.2 Proof of Lemma 4**

**Proof.** Recall that $R^* = R(h^*)$, where

$$
h^*(x) = \text{sign}(\Pr(Y = 1 \mid X = x) - 1/2).
$$

$5 \text{sign}(0) \in \{-1, 1\}$ may be defined arbitrarily without affecting the value of $R^*$. 

13
For $n \geq 3$, define
\[
A_n = \{ x \in X : \mathbb{P}(Y = 1 \mid X = x) \geq 1/2 + 1/n \}, \\
B_n = \{ x \in X : \mathbb{P}(Y = 1 \mid X = x) \leq 1/2 - 1/n \}, \\
C = \{ x \in X : \mathbb{P}(Y = 1 \mid X = x) = 1/2 \}.
\]

The doubling property and finite diameter of $X$ imply that it is totally bounded (i.e., for all $\alpha > 0$, $X$ be covered by finitely many balls of diameter $\alpha$, [13]), and hence compact. Thus, Urysohn’s lemma [20] implies that for each $n \geq 3$, there is a continuous function $f_n : X \to \mathbb{R}$ such that $f_n(A_n) = \{1\}$, $f_n(B_n) = \{-1\}$ and $f_n(C) = \{0\}$. Since continuous functions on compact sets can be uniformly approximated by Lipschitz functions, there is no loss of generality in assuming that each $f_n$ is a Lipschitz function. Normalizing by $\|f_n\|_\infty$, we have that for each $n \geq 3$, there is a $\gamma_n > 0$ and a $1$-Lipschitz function $f_n$ such that $f_n(A_n) = \{\gamma_n\}$ and $f_n(B_n) = \{-\gamma_n\}$. Since $f_n \to h^*$ pointwise, Lebesgue’s dominated convergence theorem implies that $\lim_{n \to \infty} R(f_n) = R^*$. Another application of this theorem yields that $\lim_{\gamma \to 0} R_{\gamma}(f) = R(f)$ holds for all measurable $f : X \to \mathbb{R}$.

Choosing $n$ sufficiently large that $|R(f_n) - R^*| < \varepsilon/2$ and then $\gamma$ sufficiently small that $|R_{\gamma}(f_n) - R(f_n)| < \varepsilon/2$ proves the claim. \qed

### A.3 Proof of Lemma 5

**Proof.** By rescaling $f \in \mathcal{F}_L$ to $g = f/L$ we have $g \in \mathcal{F}_1$ and
\[
\mathcal{L}_{\xi_n,1}(f) = E[\Phi_{L^{-1},L^{-1}(1+\xi_n,1)}(g)] \\
= E[\Phi_{L^{-1},L^{-1}(1+(n_d)^{-1})}(g)] \\
\leq E[\Phi_{L^{-1},L^{-1}(1+n_d^{-1})}(g)]
\]
\[
R_{\gamma=1}(f) = R_{L^{-1}}(g).
\]
So (20) is equivalent to claiming the existence of $n_0(\varepsilon) \in \mathbb{N}$ such that for all $l \in \mathbb{N}$ and $n \geq n_0(\varepsilon)$,
\[
\sup_{g \in \mathcal{F}_1} \left| \mathcal{L}_{L^{-1},L^{-1}(1+n_d^{-1})}(g) - R_{L^{-1}}(g) \right| \leq \varepsilon/4. \tag{23}
\]
Since for all $g \in \mathcal{F}_1$ we have that $\mathcal{L}_{L^{-1},L^{-1}(1+n_d\gamma)}(g)$ converges to $R_{L^{-1}}(g)$ pointwise, it follows by Lebesgue dominated convergence Theorem that
\[
\lim_{n \to \infty} \mathcal{L}_{L^{-1},L^{-1}(1+n_d\gamma)}(g) = R_{L^{-1}}(g).
\]
\qed

14