A Bayes consistent 1-NN classifier

Aryeh Kontorovich
Ben-Gurion University of the Negev

Roi Weiss
Ben-Gurion University of the Negev

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 consistent proximity-based method was the $k$-nearest neighbor classifier, for $k$ growing appropriately with sample size. We will argue that a margin-regularized 1-NN enjoys considerable statistical and algorithmic advantages over the $k$-NN classifier. 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 sophisticated techniques developed in recent years, this deceptively simple method continues to "yield[] competitive results" (Weinberger and Saul, 2009) and inspire papers in "defense of nearest-neighbor based […] classification" (Boiman et al., 2008).

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 Cover and Hart (1967), 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 risk. 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 (Stone, 1977; Devroye and Győrfi, 1985; Zhao, 1987).

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 (Hart, 1968; Gates, 1972; Ritter et al., 1975; Wilson and Martinez, 2000; Gottlieb et al., 2014b) — of which only the one in Gottlieb et al. (2014b) comes with any rigorous compression guarantees, and only for $k = 1$; moreover, it is shown therein that the condensing problem is ill-posed for $k > 1$. Query evaluation on test points may be significantly sped up via an approximate nearest neighbor search (Krauthgamer and Lee, 2004; Beygelzimer et al., 2006; Andoni and Indyk, 2008; Gottlieb et al., 2010). The price one pays for the fast approximate search is a degraded classification accuracy, and of the works cited, only Gottlieb et al. (2010) 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 von Luxburg and Bousquet (2004); Gottlieb et al. (2010, 2014b) are all for $k = 1$.

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 Gottlieb et al. (2010) and, more recently, Gottlieb et al. (2014b). 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 $\mathcal{X}$ 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 $\hat{S} \subseteq S$ such that

(i) the 1-NN classifier induced by $\hat{S}$ mislabels at most $\varepsilon n$ points in $S$ and

(ii) every pair of opposite-labeled points in $\hat{S}$ is at least $\gamma$
A Bayes consistent 1-NN classifier
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 Gottlieb et al. (2010, 2014b), 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 $\text{gen-err}$ is the generalization error of the 1-NN classifier induced by an $\varepsilon$-consistent, $\gamma$-separable $\hat{S} \subset S$, and the two terms on the right-hand side correspond roughly to sample error and hypothesis complexity. The approach proposed in Gottlieb et al. (2010, 2014b) 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 consisted of 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:

$$\mathbb{P}(h_n(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n)) \xrightarrow{n \to \infty} R^*,$$

where

$$R^* = \inf_{h : \mathcal{X} \to \{-1, 1\}} \mathbb{P}(h(X) \neq Y)$$

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 (1967), it was shown by Devroye and Gyo"{o}rfi (1985); Zhao (1987) 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 $n > n_0(\varepsilon, k)$,

$$\mathbb{P}(R(h_{k, \text{NN}}) > R^* + \varepsilon) \leq 2 \exp \left( -\frac{n \delta^2}{5184 \kappa_d^2} \right),$$

where $\kappa_d < \left( 1 + 2/\sqrt{2 - \sqrt{3}} \right)^d$ 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 Devroye et al. (1996)). Given the inherently Euclidean nature of $\kappa_d$, (2) does not seem to readily extend to more general metric spaces. It was (essentially) shown in Shalev-Shwartz and Ben-David (2014) that

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

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 $\eta(x) = \mathbb{P}(Y = 1 | X = x)$. Recently, some of the classic results on $k$-NN risk decay rates were refined by Chaudhuri and Dasgupta (2014) in an analysis that captures the interplay between the metric and the sampling distribution.

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 Snapp and Venkatesh (1998); Psaltis et al. (1994) likewise do not provide a computable finite-sample bound.

An entire chapter in Devroye et al. (1996) is devoted to condensed and edited NN rules. In the terminology of this paper, this amounts to extracting a sub-sample $\hat{S}$ and predicting via the 1-NN classifier induced by that $\hat{S}$. Assuming a certain sample compression rate and an oracle for choosing an optimal fixed-size $\hat{S}$, this scheme is shown to be weakly Bayes consistent. The generalizing power of sample compression was independently discovered by Littlesone and Warmuth (1986), and later elaborated upon by Graepel et al. (2005). In the context of NN classification, Devroye et al. (1996) list various condensing heuristics (which have no known 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 von Luxburg and Bousquet (2004), 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, Gottlieb et al. (2010) obtained bounds on the fat-shattering dimension of Lipschitz functions in doubling spaces and gave margin-based risk bounds decaying as $O(n^{-1/2})$ as opposed to $n^{-1/d}$. More recently, the existence of a margin was leveraged to give nearly optimal sample compression bounds, with corresponding generalization guarantees (Gottlieb et al., 2014b).

**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:
\[
diam(\mathcal{X}) := \sup_{x,x' \in \mathcal{X}} \rho(x,x') = 1.
\]

A function \( f : \mathcal{X} \to \mathbb{R} \) 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 \( \|f\|_{Lip} \), is the smallest \( L \) for which \( f \) is \( L \)-Lipschitz. The collection of all \( L \)-Lipschitz \( f : \mathcal{X} \to [-1,1] \) will be denoted by \( \mathcal{F}_L \). The distance between two sets \( A, B \subset \mathcal{X} \) 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 \( ddim(\mathcal{X}) := \log_2 \lambda \). A metric is doubling when its doubling dimension is finite. We will denote \( d := ddim(\mathcal{X}) < \infty \).

**Learning model.** We work in the standard agnostic learning model (Mohri et al., 2012; Shalev-Shwartz and Ben-David, 2014), 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 \( h : \mathcal{X} \to \{-1,1\} \), 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
\[
h^*(x) = \arg\max_{y \in \{-1,1\}} \mathbb{P}(Y = y \mid X = x)
\]
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 \( R(h_n) \to R^* \) almost surely.

**Sub-sample, margin, and induced 1-NN.** In a slight abuse of notation, we will blur the distinction between \( S \subset \mathcal{X} \) 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 \( \tilde{S} \subset S \) partitioned into its positively and negatively labeled subsets as \( S = S_+ \cup S_- \) is well-defined. The margin of \( \tilde{S} \), defined by
\[
\text{marg}(\tilde{S}) = \rho(\tilde{S}_+ \cup \tilde{S}_-),
\]
is the minimum distance between a pair of opposite-labeled points (see Fig. 1). A sub-sample \( \tilde{S} \) naturally induces the

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

## 3 Learning Algorithm: Regularized 1-NN

This section is 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 Gottlieb et al. (2010). Our point of departure is the connection made by von Luxburg and Bousquet (2004) between Lipschitz functions and 1-NN classifiers.

**Theorem 1** (von Luxburg and Bousquet (2004)). If \( \tilde{S} \) is a sub-sample with \( \text{marg}(\tilde{S}) \geq \gamma \), then there is an \( f \in \mathcal{F}_2 \) such that
\[
h_{\tilde{S}}(x) = \text{sign}(f(x))
\]
for all \( x \in \mathcal{X} \). More explicitly, \( f \in \mathcal{F}_2 \) 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 \( \mathcal{F}_2 \) that are Lipschitz-extensions of \( \gamma \)-separable sub-samples and will never need
A Bayes consistent 1-NN classifier

to actually calculate these explicitly; their only purpose
is to facilitate the analysis. In line with the Structural
Risk Minimization (SRM) paradigm, our learning algo-

rithm consists of minimizing the penalized margin risk,

\[ \hat{R}_{n,\gamma}^{\text{PEN}}(f) = \hat{R}_{n,\gamma}(f) + r^{\text{PEN}}(n, \gamma), \tag{5} \]

where

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

and \( c_1, \ cd \) are explicitly computable constants, the latter
depending only on \( d \). The form of the penalty term (which
is different from the penalty term in Gottlieb et al. (2010))
will be motivated by the analysis in the sequel.

This optimization is performed via two nested routines: the
inner one minimizes \( \hat{R}_{n,\gamma}^{\text{PEN}}(f) \) over \( f \in \mathcal{F}_2 \) 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 Gottlieb et al. (2010), we will give a high-
level sketch.

**Inner routine: optimizing over \( f \in \mathcal{F}_2 \).** By Theorem 1,
minimizing \( \hat{R}_{n,\gamma}^{\text{PEN}}(f) \) over \( f \in \mathcal{F}_2 \) for a fixed \( \gamma \) is equiva-
lent to seeking a \( \gamma \)-separable \( \overline{S} \subseteq S \) whose induced 1-NN
classifier \( h_{\overline{S}} \) makes the fewest mistakes on \( S \) (see Algo-

rithm 1). The algorithm invokes a minimum vertex cover
routine, which by König’s theorem is equivalent to max-
mum matching for bipartite graphs, and is computable in
randomized time \( O(n^{2.376}) \) (Mucha and Sankowski, 2004).

**Algorithm 1 minimizing \( \hat{R}_{n,\gamma}^{\text{PEN}}(f) \) over \( f \in \mathcal{F}_2 \) for fixed \( \gamma \)**

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

**Outer loop: minimizing over \( \gamma > 0 \).** Although \( \gamma \) takes
on a continuum of values, we need only consider those in-
duced by distances between opposite-labeled points in \( S \),
of which there are \( O(n^2) \). For each candidate \( \gamma \), Algo-

rithm 1 computes the optimal \( f^{*, \gamma}_n \in \mathcal{F}_2 \). Let \( \gamma^{*, \gamma}_n \) be a mini-
\mover of \( \hat{R}_{n,\gamma}^{\text{PEN}}(f^{*, \gamma}_n, \gamma) \), with corresponding \( f^{*, \gamma}_n := f^{*, \gamma}_n \):

\[
\hat{R}_{n,\gamma}^{\text{PEN}}(f^{*, \gamma}_n) = \inf_{\gamma > 0} \inf_{f \in \mathcal{F}_2} \hat{R}_{n,\gamma}^{\text{PEN}}(f) = \inf_{\gamma > 0} \hat{R}_{n,\gamma}(f^{*, \gamma}_n) = \hat{R}_{n,\gamma}(f^{*, \gamma}_n).
\]

The total runtime for computing \( \gamma^{*, \gamma}_n \) and \( f^{*, \gamma}_n \) is \( O(n^{4.376}) \),
which may be considerably sped up if one is willing to tol-
erate a small approximation factor (Gottlieb et al., 2010, 2014a).

**4 Consistency proof**

We now prove the main technical result of the paper:

**Theorem 2.** With probability one over the random sample \( S \) of size \( n \),

\[ \lim_{n \to \infty} R(f^{\gamma}_n) = R^*. \]

We will break it up into high-level steps. The basic plan is
standard: decompose the excess risk into two terms,

\[ R(f^{\gamma}_n) - R^* = \left( R(f^{\gamma}_n) - \hat{R}_{n,\gamma}^{\text{PEN}} \right) + \left( \hat{R}_{n,\gamma}^{\text{PEN}} - R^* \right) \]

and show that each decays to 0 almost surely. For con-
venience, the notation used in the proof is summarized in
Table 1. All omitted proofs are given in the Appendix.

**4.1 The term (I)**

In order to connect \( \hat{R}_{n,\gamma}^{\text{PEN}} \) and \( R(f^{\gamma}_n) \) we first need a concen-
tration bound. More specifically, since \( \hat{R}_{n,\gamma}^{\text{PEN}} \) involves the
optimal margin \( \gamma^{*, \gamma}_n \) (which is a priori unknown), we would
like to prove for each \( \gamma > 0 \) a deviation estimate on

\[ |R(f) - \hat{R}_{n,\gamma}(f)|, \]

uniformly over all \( f \in \mathcal{F}_2 \). We find it most convenient to
do this using Rademacher complexities\(^2\), 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 \( \Phi_{\gamma, \xi} \) and corresponding surrogate risk \( \mathcal{L}_{\gamma, \xi} \)
as follows.

\(^2\)An alternative, though somewhat messier route, would be to
use fat-shattering dimension, as in Gottlieb et al. (2010).
\[ R(f) = \gamma\text{-margin risk} \]

\[ \hat{R}_{n,\gamma}(f) = \text{empirical } \gamma\text{-margin risk} \]

\[ \hat{R}^\text{pen}_{n,\gamma}(f) = \text{penalized empirical } \gamma\text{-margin risk} \]

\[ \hat{R}^\text{pen}_{n,\gamma,*} = \text{optimal penalized empirical risk} \]

\[ f_{n,\gamma}^* = \text{optimal } f \in \mathcal{F}_2 \text{ for a fixed } \gamma \]

\[ \gamma_n^*, \gamma^* \] \text{ optimal margin and optimal } f \in \mathcal{F}_2

\[ \hat{\mathcal{L}}_{n,\gamma,\xi}(f) = \text{surrogate risk} \]

\[ \hat{\mathcal{L}}_{n,\gamma,\xi}(f) = \text{empirical surrogate risk} \]

**Surrogate loss.** For \( 0 < \gamma, \xi \leq 1 \) define the *surrogate loss* function \( \Phi_{\gamma,\xi}(u) : \mathbb{R} \to [0,1] \)

\[
\Phi_{\gamma,\xi}(u) = \begin{cases} 
1 & \text{if } u \leq \gamma(1 - \xi), \\
0 & \text{if } u \geq \gamma, \\
(\gamma - u)/(\gamma \xi) & \text{otherwise},
\end{cases}
\tag{9}
\]

illustrated in Figure 2, and its associated expected and empirical surrogate risks,

\[
\mathcal{L}_{\gamma,\xi}(f) = \mathbb{E}[\Phi_{\gamma,\xi}(Y f(X))],
\]

\[
\hat{\mathcal{L}}_{n,\gamma,\xi}(f) = \frac{1}{n} \sum_{i=1}^{n} \Phi_{\gamma,\xi}(Y_i f(X_i)).
\tag{10}
\]

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

\[
\gamma_{n,l} = (1 - \xi_n)^{-1}, \quad \xi_n = 1/n_d
\tag{11}
\]

\[
\varepsilon_n = \frac{2}{\gamma_{n,l} \xi_{n} n_d} + \sqrt{\frac{2c_1 \log \left( \frac{1}{\xi_n \log \frac{n}{\gamma_{n,l}}} \right)}{n}},
\tag{12}
\]

where

\[
n_d = \left( \frac{n}{c_d} \right)^{\frac{1}{2(d+1)}}.
\tag{13}
\]

This enables us to obtain a uniform deviation estimate:

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

\[
P\left( \exists l \in \mathbb{N} : \sup_{f \in \mathcal{F}_2} \left| \mathcal{L}_{\gamma_{n,l},\xi_n}(f) - \hat{\mathcal{L}}_{n,\gamma_{n,l},\xi_n}(f) \right| > \varepsilon + \varepsilon_n,l \right) \leq \frac{\pi^2}{6} \exp \left( -\frac{n \varepsilon^2}{c_1} \right).
\]

Armed with this uniform deviation bound, we proceed with the proof that the term (I) decays to zero almost surely. By Theorem 1 we may assume that \( f_{n}^* \in \mathcal{F}_2 \) is in the form of (4) with \( \gamma = \gamma_n^* \) being the optimal margin. Given \( \gamma_n^* \), let \( l_n^-, l_n^+ \in \mathbb{N} \) be the *consecutive* margin indexes in the stratification grid (11) such that

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

and abbreviate \( \gamma_n^+ = \gamma_{n,l_n^+} \) and \( \gamma_n^- = \gamma_{n,l_n^-} \). We now relate the margin risks to the surrogate risks. Note that since \( 0 \leq \gamma_n^- \leq \gamma_n^* \leq \gamma_n^+ \), we have

\[
R(f_n^*) \leq \mathcal{L}_{\gamma_n^-\xi_n}(f_n^*),
\]

\[
\hat{R}_{n,\gamma_n^-}(f_n^*) \geq \hat{\mathcal{L}}_{n,\gamma_n^-\xi_n}(f_n^*),
\]

\[
r_{\text{pen}}(n, \gamma_n^*) \geq r_{\text{pen}}(n, \gamma_n^+).
\]

Thus,

\[
(*) \quad := \quad P\left( R(f_n^*) - \hat{R}^\text{pen}_{n,\gamma} > \varepsilon \right)
\]

\[
= \quad P\left( R(f_n^*) - \hat{R}_{n,\gamma_n^-}(f_n^*) - r_{\text{pen}}(n, \gamma_n^*) > \varepsilon \right)
\leq \quad P\left( \mathcal{L}_{\gamma_n^-\xi_n}(f_n^*) - \hat{\mathcal{L}}_{n,\gamma_n^-\xi_n}(f_n^*) > \varepsilon + r_{\text{pen}}(n, \gamma_n^-) \right),
\]

and since \( f_{n}^* \in \mathcal{F}_2 \), we have

\[
(*) \quad \leq \quad P\left( \sup_{f \in \mathcal{F}_2} \left| \mathcal{L}_{\gamma_n^-\xi_n}(f) - \hat{\mathcal{L}}_{n,\gamma_n^-\xi_n}(f) \right| > \varepsilon + \varepsilon_n,l \right)
\]

Next, we make a connection between \( r_{\text{pen}}(n, \gamma_{n,l-1}) \) and \( \varepsilon_n,l \), justifying the form of the penalty term in (6):

**Lemma 4.** For all \( l \geq 2 \) and all \( n \) sufficiently large,

\[
r_{\text{pen}}(n, \gamma_{n,l-1}) \geq \varepsilon_n,l.
\]
An application of Lemma 4 yields

$$(*) \leq \mathbb{P}\left( \exists l \geq 2 : \sup_{f \in \mathcal{F}_2} |\mathcal{L}_{\gamma_n,l,\xi_n}(f) - \hat{\mathcal{L}}_{\gamma_n,l,\xi_n}(f)| > \varepsilon + \varepsilon_n,l \right)$$

$$\leq \frac{\pi^2}{6} \exp\left( -\frac{n\varepsilon^2}{c_1} \right),$$

where the last inequality follows from Lemma 3.

### 4.2 The term (II)

We begin by approximating the Bayes optimal risk by the margin risk:

**Lemma 5.** For every $\varepsilon > 0$ there is a $\gamma > 0$ such that

$$\inf_{f \in \mathcal{F}_2} R_\gamma(f) - R^* < \varepsilon.$$  

In particular,

$$R^* = \lim_{\gamma \to 0} \inf_{f \in \mathcal{F}_2} R_\gamma(f). \quad (14)$$

Since (14) holds for any sequence $\gamma_n \rightarrow 0$, it is true in particular of subsequences of the stratification grid (11). Hence, for all $\varepsilon > 0$, there is a $\hat{\gamma}^+$ with a corresponding $\hat{f}^+ \in \mathcal{F}_2$ such that

$$\inf_{f \in \mathcal{F}_2} R_{\hat{\gamma}^+}(f) \leq R^* + \varepsilon/8,$$

$$R_{\hat{\gamma}^+}(\hat{f}^+) \leq \inf_{f \in \mathcal{F}_2} R_{\hat{\gamma}^+}(f) + \varepsilon/8.$$  

Fix such a $\hat{\gamma}^+$ and let $\hat{\gamma}^-$ be the “next” margin in the stratification (11). Now by (7), Algorithm 1 provides an optimal $f^*_n$ such that

$$\hat{R}_{\gamma_n}^{\text{Pen}}(f^*_n) \leq \hat{R}_{\gamma_n}^{\text{Pen}}(f^*_{\gamma_n}) \leq \hat{R}_{\gamma_n}^{\text{Pen}}(\hat{f}^+).$$

Hence, for the term (II) we have

$$(**) := \mathbb{P}\left( \hat{R}_{\gamma_n}^{\text{Pen}} - R^* > \varepsilon \right)$$

$$\leq \mathbb{P}\left( \hat{R}_{\gamma_n}^{\text{Pen}} - R_{\hat{\gamma}^+}(\hat{f}^+) > 3\varepsilon/4 \right)$$

$$\leq \mathbb{P}\left( \hat{R}_{\gamma_n}^{\text{Pen}}(\hat{f}^+) - R_{\hat{\gamma}^+}(\hat{f}^+) > 3\varepsilon/4 - n^{\text{Pen}}(n, \hat{\gamma}^-) \right).$$

Next, note that the margin loss $R_{\hat{\gamma}^+}(\cdot)$ is well-approximated by surrogate losses:

**Lemma 6.** For every $\gamma > 0$ and $f \in \mathcal{F}_2$

$$\lim_{n \to \infty} |\mathcal{L}_{\gamma,\xi_n}(f) - R_\gamma(f)| = 0. \quad (15)$$

Hence, we may take $n$ sufficiently large so that

$$|\mathcal{L}_{\hat{\gamma}^+,\xi_n}(\hat{f}^+) - R_{\hat{\gamma}^+}(\hat{f}^+)| \leq \varepsilon/4.$$

Since by construction,

$$\gamma_n,l+1 \gamma_n,l = 1 - \xi_n, \quad \forall l \in \mathbb{N},$$

it follows that $\hat{\gamma}^- = \hat{\gamma}^+(1 - \xi_n)$ and thus

$$\hat{R}_{\gamma_n}^{\text{Pen}}(\hat{f}^+) \leq \hat{R}_{\gamma_n}^{\text{Pen}}(\hat{f}^+).$$

Taking $n$ sufficiently large to ensure $n^{\text{Pen}}(n, \hat{\gamma}^-) \leq \varepsilon/4$ and combining these estimates yields

$$(**) \leq \mathbb{P}\left( \hat{R}_{\gamma_n}^{\text{Pen}}(\hat{f}^+) - R_{\hat{\gamma}^+,\xi_n}(\hat{f}^+) > \varepsilon/4 \right)$$

$$\leq \mathbb{P}\left( \sup_{f \in \mathcal{F}_2} \hat{\mathcal{L}}_{\gamma_n,\xi_n}(f) - \mathcal{L}_{\hat{\gamma}^+,\xi_n}(f) > \varepsilon/4 \right)$$

$$\leq c e^{-\frac{n\varepsilon^2}{c_1}},$$

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 $\rho(x, x') = ||x - x'||_2$, and defined a joint distribution over $\mathcal{X} \times \{-1, 1\}$ as follows. A point $(x_1, x_2) \in \mathbb{R}^2$ is sampled by drawing $T \in [0, 2\pi]$ uniformly at random and putting

$$x_1(T) = A\sqrt{T} \cos(\omega T),$$

$$x_2(T) = A\sqrt{T} \sin(\omega T)$$

Figure 3: The distribution for $A = 5$ and $\omega = 3$.  

485
for some specified parameters $A$ and $\omega$. The label $Y \in \{-1, 1\}$ is drawn according to the conditional distribution

$$\eta(T) = P(Y = 1 | T) = \frac{1 + \cos(\omega T)}{2},$$

as illustrated in Figure 3.

We compared four classifiers: $k^*$-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 while searching for the optimal margin). Their runtime and generalization performance, averaged over 100 independent runs, are summarized in Figures 4 and 5.

Our proposed algorithm, SRM-1-NN, emerges competitive by both criteria.

A Appendix

A.1 Proof of Lemma 3

We first need the following uniform convergence lemma.

**Lemma 7.** For any $0 < \varepsilon, 0 < \xi < 1$ and $0 < \gamma$,

$$P\left(\sup_{f \in F_2} \left| L_{\gamma, \xi}(f) - \hat{L}_{n, \gamma, \xi}(f) \right| > 2R_n(L_{\gamma, \xi} \circ F_2) + \varepsilon\right) \leq \exp\left(-n\varepsilon^2/c_1\right),$$

where the Rademacher complexity $R_n(L_{\gamma, \xi} \circ F_2)$ satisfies

$$R_n(L_{\gamma, \xi} \circ F_2) \leq \frac{2}{\gamma^2} \left(\frac{c_1}{n}\right)^{1/\xi} =: \hat{R}_{n, \gamma, \xi}. \tag{17}$$

**Proof of Lemma 7.** Equation (16) is restatement of Mohri et al. (2012, Theorem 3.1). Note that $\Phi_{\gamma, \xi} : \mathbb{R} \to [0, 1]$ is $1/(\gamma \xi)$-Lipschitz. Thus, by Talagrand’s contraction lemma (Ledoux and Talagrand, 1991),

$$R_n(L_{\gamma, \xi} \circ F_2) \leq \frac{2}{\gamma^2} \frac{c_1}{n} =: \hat{R}_{n, \gamma, \xi}.$$  \tag{17}

The upper estimate on $R_n(F_1)$ implicit in (17) is essentially contained in Equation (10) of Kontorovich and Weiss (2014).

**Proof of Lemma 3.** Following proof idea in Devroye et al.
A Bayes consistent 1-NN classifier

(1996, Theorem 18.2), a union bound yields

\[ (***) \colon= P\left( \exists l \in \mathbb{N} : \left| L_{\gamma_{n,l},\xi_n}(f) - \hat{L}_{\gamma_{n,l},\xi_n}(f) \right| > \varepsilon + \varepsilon_{n,l} \right) \leq \sum_{l=1}^{\infty} P\left( \sup_{f \in \mathcal{F}_2} \left| L_{\gamma_{n,l},\xi_n}(f) - \hat{L}_{\gamma_{n,l},\xi_n}(f) \right| > \varepsilon + \varepsilon_{n,l} \right). \]

Note that by construction

\[ \varepsilon_{n,l} = 2R_{n,\gamma_{n,l},\xi_n} + \sqrt{\frac{2c_1 \log \left( \frac{1}{\xi_n} \log \frac{2}{\gamma_{n,l}} \right)}{n}}. \]

Thus, writing

\[ r_{n,l} = \sqrt{\frac{2c_1 \log \left( \frac{1}{\xi_n} \log \frac{2}{\gamma_{n,l}} \right)}{n}} \]

and bounding each term in the sum by Lemma 7, we have

\[ (***) \leq \sum_{l=1}^{\infty} \exp \left( - \frac{n(\varepsilon + r_{n,l})^2}{c_1} \right) \leq \exp \left( - \frac{n\varepsilon^2}{c_1} \right) \sum_{l=1}^{\infty} \exp \left( - \frac{n r_{n,l}^2}{c_1} \right). \]

Next note that by the definition of \( \gamma_{n,l} \) we have

\[ \gamma_{n,l} = (1 - \xi_n)^{-1} \leq \exp (-\xi_n(l - 1)). \]

Solving for \( l \) yields

\[ \frac{1}{\xi_n} \log \frac{e}{\gamma_{n,l}} \geq l. \]

Thus,

\[ \exp \left( - \frac{n r_{n,l}^2}{c_1} \right) \leq \frac{1}{l^2} \]

and summing over \( l \) yields the claim. \( \square \)

A.2 Proof of Lemma 4

Let us first write \( \varepsilon_{n,l} \) in terms of \( \gamma_{n,l-1} \). Since \( \gamma_{n,l} = \gamma_{n,l-1}(1 - \xi_n) \) by definition, we have

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

Taking \( n \) sufficiently large to ensure that \( 1 - \xi_n \geq 1/2 \) we have that for all \( l \geq 2, \)

\[ \varepsilon_{n,l} \leq \frac{4}{\gamma_{n,l-1}(1 - \xi_n) n_d} + \sqrt{2c_1 \log \left( \frac{1}{\xi_n} \log \frac{2}{\gamma_{n,l-1}} \right) n}, \]

which is exactly \( R_{\text{pen}}(n,\gamma_{n,l-1}) \).

A.3 Proof of Lemma 5

Recall that \( R^* = R(h^*) \), where

\[ h^*(x) = \text{sign}(P(Y = 1 | X = x) - 1/2). \]

For \( n \geq 3 \), define

\[ A_n = \{ x \in \mathcal{X} : P(Y = 1 | X = x) \geq 1/2 + 1/n \}, \]

\[ B_n = \{ x \in \mathcal{X} : P(Y = 1 | X = x) \leq 1/2 - 1/n \}, \]

\[ C = \{ x \in \mathcal{X} : P(Y = 1 | X = x) = 1/2 \}. \]

The doubling property and finite diameter of \( \mathcal{X} \) imply that it is totally bounded (i.e., for all \( \alpha > 0 \), \( \mathcal{X} \) can be covered by finitely many balls of diameter \( \alpha \) (Krauthgamer and Lee, 2004)), and hence compact. Thus, Urysohn’s lemma (Rudin, 1987) implies that for each \( n \geq 3 \), there is a continuous \( f_n : \mathcal{X} \to \mathbb{R} \) 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 \( \|f_n\|_{\text{lip}} \), we have that for each \( n \geq 3 \), there is a \( \gamma_n > 0 \) and a 1-Lipschitz \( f_n \) 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 \( f : \mathcal{X} \to \mathbb{R} \).

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

A.4 Proof of Lemma 6

Rescaling \( f \in \mathcal{F}_2 \) to \( g = 2f/\gamma \), Eq. (15) is equivalent to claiming the existence of an \( n_0(\varepsilon) \in \mathbb{N} \) such that for all \( n \geq n_0(\varepsilon) \),

\[ |L_{1,\xi_n}(g) - R_1(g)| \leq \varepsilon/4. \]

Since \( \xi_n = n_d^{-1} \) decays to zero with increasing \( n \), it follows that \( L_{1,n_d^{-1}}(g) \to R_1(g) \) pointwise, and so by Lebesgue’s dominated convergence theorem, we have that

\[ \lim_{n \to \infty} L_{1,n_d^{-1}}(g) = R_1(g), \]

proving the claim.

\( \square \)
References


