An adaptive nearest neighbor rule for classification

Akshay Balsubramani
abalsubr@stanford.edu

Sanjoy Dasgupta
dasgupta@eng.ucsd.edu

Yoav Freund
yfreund@eng.ucsd.edu

Shay Moran
shaym@princeton.edu

Abstract

We introduce a variant of the k-nearest neighbor classifier in which k is chosen adaptively for each query, rather than being supplied as a parameter. The choice of k depends on properties of each neighborhood, and therefore may significantly vary between different points. For example, the algorithm will use larger k for predicting the labels of points in noisy regions.

We provide theory and experiments that demonstrate that the algorithm performs comparably to, and sometimes better than, k-NN with an optimal choice of k. In particular, we bound the convergence rate of our classifier in terms of a local quantity we call the "advantage", giving results that are both more general and more accurate than the smoothness-based bounds of earlier nearest neighbor work. Our analysis uses a variant of the uniform convergence theorem of Vapnik-Chervonenkis that is for empirical estimates of conditional probabilities and may be of independent interest.

1 Introduction

We introduce an adaptive nearest neighbor classification rule. Given a training set with labels $\{\pm 1\}$, its prediction at a query point x is based on the training points closest to x, rather like the k-nearest neighbor rule. However, the value of k that it uses can vary from query to query. Specifically, if there are n training points, then for any query x, the smallest k is sought for which the k points closest to x have labels whose average is either greater than $+\Delta(n,k)$, in which case the prediction is $+1$, or less than $-\Delta(n,k)$, in which case the prediction is -1; and if no such k exists, then "?" ("don't know") is returned. Here, $\Delta(n,k) \sim \sqrt{(\log n)/k}$ corresponds to a confidence interval for the average label in the region around the query.

We study this rule in the standard statistical framework in which all data are i.i.d. draws from some unknown underlying distribution P on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the data space and $\mathcal{Y}$ is the label space. We take $\mathcal{X}$ to be a separable metric space, with distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and we take $\mathcal{Y} = \{\pm 1\}$. We can decompose P into the marginal distribution μ on $\mathcal{X}$ and the conditional expectation of the label at each point x: if (X,Y) represents a random draw from P, define $\eta(x) = \mathbb{E}(Y|X=x)$. In this terminology, the Bayes-optimal classifier is the rule $g^* : \mathcal{X} \to \{\pm 1\}$ given by

$$g^*(x) = \begin{cases} \text{sign}(\eta(x)) & \text{if } \eta(x) \neq 0 \\ \text{either } -1 \text{ or } +1 & \text{if } \eta(x) = 0 \end{cases} \tag{1}$$

and its error rate is the Bayes risk, $R^* = \frac{1}{2}\mathbb{E}_{X\sim\mu}\left[1 - |\eta(X)|\right]$. A variety of nonparametric classification schemes are known to have error rates that converge asymptotically to R^*. These include k-nearest neighbor (henceforth, k-NN) rules [FH51] in which k grows with the number of training points n according to a suitable schedule (k_n), under certain technical conditions on the metric measure space $(\mathcal{X}, d, \mu)$.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

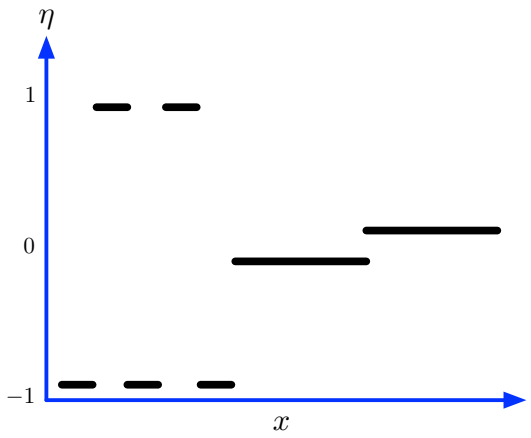


Figure 1: For values of x on the left half of the shown interval, the pointwise bias $\eta(x)$ is close to -1 or 1, and thus a small value of k will yield an accurate prediction. Larger k will not do as well, because they may run into neighboring regions with different labels. For values of x on the right half of the interval, $\eta(x)$ is close to 0, and thus large k is essential for accurate prediction.

In this paper, we are interested in consistency as well as rates of convergence. In particular, we find that the adaptive nearest neighbor rule is also asymptotically consistent (under the same technical conditions) while converging at a rate that is about as good as, and sometimes significantly better than, that of k-NN under any schedule (k_n).

Intuitively, one of the advantages of k-NN over nonparametric classifiers that use a fixed bandwidth or radius, such as Parzen window or kernel density estimators, is that k-NN automatically adapts to variation in the marginal distribution μ: in regions with large μ, the k nearest neighbors lie close to the query point, while in regions with small μ, the k nearest neighbors can be further afield. The adaptive NN rule that we propose goes further: it also adapts to variation in η. In certain regions of the input space, where η is close to 0, an accurate prediction would need large k. In other regions, where η is near -1 or 1, a small k would suffice, and in fact, a larger k might be detrimental because neighboring regions might be labeled differently. See Figure 1 for one such example. A k-NN classifier is forced to pick a single value of k that trades off between these two contingencies. Our adaptive NN rule, however, can pick the right k in each neighborhood separately.

Our estimator allows us to give rates of convergence that are tighter and more transparent than those customarily obtained in nonparametric statistics. Specifically, for any point x in the instance space $\mathcal{X}$, we define a notion of the *advantage at* x, denoted $\mathrm{adv}(x)$, which is rather like a local margin. We show that the prediction at x is very likely to be correct once the number of training points exceeds $\tilde{O}(1/\mathrm{adv}(x))$. Universal consistency follows by establishing that almost all points have positive advantage.

1.1 Relation to other work in nonparametric estimation

For linear separators and many other *parametric* families of classifiers, it is possible to give rates of convergence that hold without any assumptions on the input distribution μ or the conditional expectation function η. This is not true of nonparametric estimation: although any target function can in principle be captured, the number of samples needed to achieve a specific level of accuracy will inevitably depend upon aspects of this function such as how fast it changes [DGL96, chapter 7]. As a result, nonparametric statistical theory has focused on (1) asymptotic consistency, ideally without assumptions, and (2) rates of convergence under a variety of smoothness assumptions.

Asymptotic consistency has been studied in great detail for the k-NN classifier, when k is allowed to grow with the number of data points n. The risk of the classifier, denoted R_n, is its error rate on the underlying distribution P; this is a random variable that depends upon the set of training points seen. Cover and Hart [CH67] showed that in general metric spaces, under the assumption that every x in the support of μ is either a continuity point of η or has $\mu(\{x\}) > 0$, the expected risk $\mathbb{E}R_n$ converges to the Bayes-optimal risk R^*, as long as $k \to \infty$ and $k/n \to 0$. For points

in finite-dimensional Euclidean space, a series of results starting with Stone [Sto77] established consistency without any assumptions on μ or η, and showed that $R_n \to R^*$ almost surely [DGKL94]. More recent work has extended these *universal consistency* results—that is, consistency without assumptions on η—to arbitrary metric measure spaces $(\mathcal{X}, d, \mu)$ that satisfy a certain differentiation condition [CG06, CD14].

Rates of convergence have been obtained for k-nearest neighbor classification under various smoothness conditions including Holder conditions on η [KP95, Gyö81] and "Tsybakov margin" conditions [MT99, AT07, CD14]. Such assumptions have become customary in nonparametric statistics, but they leave a lot to be desired. First, they are uncheckable: it is not possible to empirically determine the smoothness given samples. Second, they view the underlying distribution P through the tiny window of two or three parameters, obscuring almost all the remaining structure of the distribution that also influences the rate of convergence. Finally, because nonparametric estimation is often *local*, there is the intriguing possibility of getting different rates of convergence in different regions of the input space: a possibility that is immediately defeated by reducing the entire space to two smoothness constants.

The first two of these issues are partially addressed by the work of [CD14], who analyze the finite sample risk of k-NN classification without any assumptions on P. Their bounds involve terms that measure the probability mass of the input space in a carefully defined region around the decision boundary: that is, bounds that are tailored to the specific distribution P, rather than reflecting worst-case behavior over some large class to which P belongs. However, the expressions for the risk are somewhat hard to parse, in large part because of the interaction between n and k.

In the present paper, we obtain finite-sample rates of convergence that are fine-tuned not just to the specific distribution P but also to the specific query point. This is achieved by defining a *margin*, or *advantage*, at every point in the input space, and giving bounds (Theorem 1) entirely in terms of this quantity. For parametric classification, it has become common to define a notion of margin that controls generalization. In the nonparametric setting, it makes sense that the margin would in fact be a function $\mathcal{X} \to \mathbb{R}$, and would yield different generalization error bounds in different regions of space. Our adaptive nearest neighbor classifier allows us to realize this vision in a fairly elementary manner.

The advantages of setting k locally have been pointed out and quantified in recent work on nonparametric *regression* [DGKL94, CS18], notably that of [Kpo11]. Although it is common to reduce classification to regression in nonparametric analysis, the right choice of k may be fundamentally different in the two settings. This is reflected in the difference between our setting for k and that of [Kpo11]; for instance, the physical value of the radius containing k points matters in that work while playing no role in ours. Moreover, the benefit of local adaptivity may be more pronounced for classification than for regression. Our analysis shows, for instance, that there is a radius r_x around each point x such that prediction based on training points in $B(x, r_x)$ will with high probability be perfect, provided there are enough such points. This is not true of regression, where the target y is a real value and thus the radius needs to keep shrinking.

Organization. Most proofs are relegated to the appendices.

In Section 2, we introduce the formal model of learning and define some basic geometric notions, as a prelude to presenting the adaptive k-NN algorithm in Section 3. In Sections 4 and 5 and Appendix A, we state and prove consistency and generalization bounds for this classifier, and compare them with prior work in the k-NN literature. Our bounds exploit a general VC-based uniform convergence statement which is presented in Section 6 and proved in a self-contained manner in Appendix B.

2 Setup

Take the instance space to be a separable metric space $(\mathcal{X}, d)$ and the label space to be $\mathcal{Y} = \{\pm 1\}$. All data are assumed to be drawn i.i.d. from a fixed unknown distribution P over $\mathcal{X} \times \mathcal{Y}$.

Let μ denote the marginal distribution on $\mathcal{X}$: if (X, Y) is a random draw from P, then

$$\mu(S) = \Pr(X \in S)$$

for any measurable set $S \subseteq \mathcal{X}$. For any $x \in \mathcal{X}$, the conditional expectation, or *bias*, of Y given x, is

$$\eta(x) = \mathbb{E}(Y|X = x) \in [-1, 1].$$

Similarly, for any measurable set S with $\mu(S) > 0$, the conditional expectation of Y given $X \in S$ is

$$\eta(S) = \mathbb{E}(Y|X \in S) = \frac{1}{\mu(S)} \int_S \eta(x)\, d\mu(x).$$

The risk of a classifier $g : \mathcal{X} \to \{-1, +1, ?\}$ is the probability that it is incorrect on pairs $(X, Y) \sim P$,

$$R(g) = P(\{(x, y) : g(x) \neq y\}). \tag{2}$$

The Bayes-optimal classifier g^*, as given in (1), depends only on η, but its risk R^* depends on μ. For a classifier g_n based on n training points from P, we will be interested in whether $R(g_n)$ converges to R^*, and the rate at which this convergence occurs.

The algorithm and analysis in this paper depend heavily on the probability masses and biases of balls in $\mathcal{X}$. For $x \in \mathcal{X}$ and $r \geq 0$, let $B(x, r)$ denote the closed ball of radius r centered at x,

$$B(x, r) = \{z \in \mathcal{X} : d(x, z) \leq r\}.$$

For $0 \leq p \leq 1$, let $r_p(x)$ be the smallest radius r such that $B(x, r)$ has probability mass at least p, that is,

$$r_p(x) = \inf\{r \geq 0 : \mu(B(x, r)) \geq p\}. \tag{3}$$

It follows that $\mu(B(x, r_p(x))) \geq p$.

The *support* of the marginal distribution μ plays an important role in convergence proofs and is formally defined as

$$\text{supp}(\mu) = \{x \in \mathcal{X} : \mu(B(x, r)) > 0 \text{ for all } r > 0\}.$$

It is a well-known consequence of the separability of $\mathcal{X}$ that $\mu(\text{supp}(\mu)) = 1$ [CH67].

3 The adaptive k-nearest neighbor algorithm

The algorithm is given a labeled training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$. Based on these points, it is able to compute empirical estimates of the probabilities and biases of different balls.

For any set $S \subseteq \mathcal{X}$, we define its empirical count and probability mass as

$$\begin{aligned} \#_n(S) &= |\{i : x_i \in S\}| \\ \mu_n(S) &= \frac{\#_n(S)}{n}. \end{aligned} \tag{4}$$

If this is non-zero, we take the empirical bias to be

$$\eta_n(S) = \frac{\sum_{i:x_i \in S} y_i}{\#_n(S)}. \tag{5}$$

The adaptive k-NN algorithm (AKNN) is shown in Figure 2. It makes a prediction at x by growing a ball around x until the ball has significant bias, and then choosing the corresponding label. In some cases, a ball of sufficient bias may never be obtained, in which event "?" is returned. In what follows, let $g_n : \mathcal{X} \to \{-1, +1, ?\}$ denote the AKNN classifier.

Later, we will also discuss a variant of this algorithm in which a modified confidence interval,

$$\Delta(n, k, \delta) = c_1 \sqrt{\frac{d_0 \log n + \log(1/\delta)}{k}} \tag{7}$$

is used, where d_0 is the VC dimension of the family of balls in $(\mathcal{X}, d)$.

In comparing the algorithm of Figure 2 to standard k-nearest neighbor classification, it might at first glance seem that we have merely replaced one parameter (k) with another (δ). This is not accurate. Our δ is the customary confidence parameter of statistics and learning theory: it provides an upper bound on the failure probability of the algorithm. It can be set to 0.05, for instance. The algorithm makes infinitely many parameter choices—it sets k for each query point—and asks for just a single failure probability that lets it know how aggressively to set its confidence intervals.

Given:

- training set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \{\pm 1\}$
- confidence parameter $0 < \delta < 1$

To predict at $x \in \mathcal{X}$:

- For any integer k, let $B_k(x)$ denote the smallest ball centered at x that contains exactly k training points. [a]
- Find the smallest $0 < k \leq n$ for which the $B_k(x)$ has a *significant bias*: that is, $|\eta_n(B_k(x))| > \Delta(n, k, \delta)$, where

$$\Delta(n, k, \delta) = c_1 \sqrt{\frac{\log n + \log(1/\delta)}{k}}. \tag{6}$$

- If there exists such a ball, return label $\text{sign}(\eta_n(B_k(x)))$.
- If no such ball exists: return "?"

[a] When several points have the same distance to x, there might be some values of k for which $B_k(x)$ is undefined. Our algorithm skips such values of k.

Figure 2: The adaptive k-NN (AKNN) classifier. The absolute constant c_1 is from Lemma 7.

4 Pointwise advantage and rates of convergence

We now provide finite-sample rates of convergence for the adaptive nearest neighbor rule. For simplicity, we give convergence rates that are specific to any query point x and that depend on a suitable notion of the "margin" of distribution P around x.

Pick any $p, \gamma > 0$. Recalling definition (3), we say a point $x \in \mathcal{X}$ is (p, γ)-salient if the following holds for either $s = +1$ or $s = -1$:

- $s\eta(x) > 0$, and $s\eta(B(x, r)) > 0$ for all $r \in [0, r_p(x))$, and $s\eta(B(x, r_p(x))) \geq \gamma$.

In words, this means that $g^*(x) = s$ (recall that g^* is the Bayes classifier), that the biases of all balls of radius $\leq r_p(x)$ around x have the same sign as s, and that the bias of the ball of radius $r_p(x)$ has absolute value at least γ. A point x can satisfy this definition for a variety of pairs (p, γ). The *advantage* of x is taken to be the largest value of $p\gamma^2$ over all such pairs: wWe will see (Lemma 3) that under a mild condition on the underlying metric measure space, almost all x with $\eta(x) \neq 0$ have a positive advantage.

4.1 Advantage-based finite-sample bounds

We now state two generalization bounds for the adaptive nearest neighbor classifier. The first holds pointwise—it bounds the probability of error at a specific point x—while the second is the type of uniform convergence bound that is more standard in learning theory.

The following theorem shows that for every point x, if the sample size n satisfies $n \gtrapprox 1/\text{adv}(x)$, then the label of x is likely to be $g^*(x)$, where g^* is the Bayes optimal classifier. This provides pointwise convergence of $g(x)$ to $g^*(x)$ at a rate that is sensitive to the local geometry of x.

Theorem 1 (Pointwise convergence rate)**.** *There is an absolute constant $C > 0$ for which the following holds. Let $0 < \delta < 1$ denote the confidence parameter in the* AKNN *algorithm (Figure 2), and suppose the algorithm is used to define a classifier g_n based on n training points chosen i.i.d. from P. Then, for every point $x \in \text{supp}(\mu)$, if*

$$n \geq \frac{C}{\text{adv}(x)} \max\left(\log \frac{1}{\text{adv}(x)},\ \log \frac{1}{\delta}\right)$$

then with probability at least $1 - \delta$ we have that $g_n(x) = g^*(x)$.

If we further assume that the family of all balls in the space has finite VC dimension d_0 then we can strengthen the guarantee to hold with high probability *simultaneously* for all $x \in \text{supp}(\mu)$. This is achieved by a modified version of the algorithm that uses confidence interval (7) instead of (6).

Theorem 2 (Uniform convergence rate)**.** *Suppose that the set of balls in $(\mathcal{X}, d)$ has finite VC dimension d_0, and that the algorithm of Figure 2 uses confidence interval (7) instead of (6). Then, with probability at least $1 - \delta$, the resulting classifier g_n satisfies the following: for every point $x \in \text{supp}(\mu)$, if*

$$n \geq \frac{C}{\text{adv}(x)} \max\left(\log \frac{1}{\text{adv}(x)},\, \log \frac{1}{\delta}\right)$$

then $g_n(x) = g^*(x)$.

A key step towards proving Theorems 1 and 2 is to identify the subset of $\mathcal{X}$ that is likely to be correctly classified for a given number of training points n. This follows the rough outline of [CD14], which gave rates of convergence for k-nearest neighbor, but there are two notable differences. First, we will see that the likely-correct sets obtained in that earlier work (for k-NN) are, roughly, subsets of those we obtain for the new adaptive nearest neighbor procedure. Second, the proof for our setting is considerably more streamlined; for instance, there is no need to devise tie-breaking strategies for deciding the identities of the k nearest neighbors.

4.2 A comparison with k-nearest neighbor

For $a \geq 0$, let $\mathcal{X}_a$ denote all points with advantage greater than a:

$$\mathcal{X}_a = \{x \in \text{supp}(\mu) : \text{adv}(x) > a\}. \tag{8}$$

In particular, $\mathcal{X}_0$ consists of all points with positive advantage.

By Theorem 1, points in $\mathcal{X}_a$ are likely to be correctly classified when the number of training points is $\widetilde{\Omega}(1/a)$, where the $\widetilde{\Omega}(\cdot)$ notation ignores logarithmic terms. In contrast, the work of [CD14] showed that with n training points, the k-NN classifier is likely to correctly classify the following set of points:

$$\begin{aligned}\mathcal{X}'_{n,k} = &\{x \in \text{supp}(\mu) : \eta(x) > 0, \eta(B(x,r)) \geq k^{-1/2} \text{ for all } 0 \leq r \leq r_{k/n}(x)\} \\ &\cup \{x \in \text{supp}(\mu) : \eta(x) < 0, \eta(B(x,r)) \leq -k^{-1/2} \text{ for all } 0 \leq r \leq r_{k/n}(x)\}.\end{aligned}$$

Such points are $(k/n, k^{-1/2})$-salient and thus have advantage at least $1/n$. In fact,

$$\bigcup_{1 \leq k \leq n} \mathcal{X}'_{n,k} \subseteq \mathcal{X}_{1/n}.$$

In this sense, the adaptive nearest neighbor procedure is able to perform *roughly* as well as all choices of k simultaneously. This is not a precise statement because of logarithmic factors (the sample complexity in Theorem 1 is $(1/a)\log(1/a)$ rather than $1/a$), and the resulting gap can be seen in our experiments.

5 Universal consistency

In this section we study the convergence of $R(g_n)$ to the Bayes risk R^* as the number of points n grows. An estimator is described as *universally consistent* in a metric measure space $(\mathcal{X}, d, \mu)$ if it has this desired limiting behavior for all conditional expectation functions η.

Earlier work [CD14] established the universal consistency of k-nearest neighbor (for $k/n \to 0$ and $k/(\log n) \to \infty$) in any metric measure space that satisfies the Lebesgue differentiation condition: that is, for any bounded measurable $f : \mathcal{X} \to \mathbb{R}$ and for almost all (μ-a.e.) $x \in \mathcal{X}$,

$$\lim_{r \downarrow 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} f \, d\mu = f(x). \tag{9}$$

This is known to hold, for instance, in any finite-dimensional normed space or any doubling metric space [Hei01, Chapter 1].

We will now see that this same condition implies the universal consistency of the adaptive nearest neighbor rule. To begin with, it implies that almost every point has a positive advantage.

Lemma 3. *Suppose metric measure space $(\mathcal{X}, d, \mu)$ satisfies condition (9). Then, for any conditional expectation η, the set of points*

$$\{x \in \mathcal{X} : \eta(x) \neq 0,\ \mathrm{adv}(x) = 0\}$$

has zero μ-measure.

Proof. Let $\mathcal{X}' \subseteq \mathcal{X}$ consist of all points $x \in \mathrm{supp}(\mu)$ for which condition (9) holds true with $f = \eta$, that is, $\lim_{r \downarrow 0} \eta(B(x,r)) = \eta(x)$. Since $\mu(\mathrm{supp}(\mu)) = 1$, it follows that $\mu(\mathcal{X}') = 1$.

Pick any $x \in \mathcal{X}'$ with $\eta(x) \neq 0$; without loss of generality, $\eta(x) > 0$. By (9), there exists $r_o > 0$ such that

$$\eta(B(x,r)) \geq \eta(x)/2 \text{ for all } 0 \leq r \leq r_o.$$

Thus x is (p, γ)-salient for $p = \mu(B(x, r_o)) > 0$ and $\gamma = \eta(x)/2$, and has positive advantage. □

Universal consistency follows as a consequence; the proof details are deferred to Appendix A.

Theorem 4 (Universal consistency)**.** *Suppose the metric measure space $(\mathcal{X}, d, \mu)$ satisfies condition (9). Let (δ_n) be a sequence in $[0,1]$ with (1) $\sum_n \delta_n < \infty$ and (2) $\lim_{n\to\infty}(\log(1/\delta_n))/n = 0$. Let the classifier $g_{n,\delta_n} : \mathcal{X} \to \{-1, +1, ?\}$ be the result of applying the* AKNN *procedure (Figure 2) with n points chosen i.i.d. from P and with confidence parameter δ_n. Letting $R_n = R(g_{n,\delta_n})$ denote the risk of g_{n,δ_n}, we have $R_n \to R^*$ almost surely.*

6 Uniform convergence of empirical conditional measures

A key piece of our analysis is a uniform convergence bound for empirical estimates of *conditional* probabilities. We now discuss this bound in an abstract setting; further details are in Appendix B.

Let P be a distribution over some space X, and let $\mathcal{A}, \mathcal{B}$ be two collections of events. Let $x_1, \ldots, x_n$ be independent samples from P. We would like to use these to estimate $P(A|B)$ simultaneously for all $A \in \mathcal{A}, B \in \mathcal{B}$. It is natural to consider the empirical estimates:

$$P_n(A|B) = \frac{\sum_i 1_{[x_i \in A \cap B]}}{\sum_i 1_{[x_i \in B]}}.$$

We study the approximation error of these estimates. Note that the case where $\mathcal{B} = \{X\}$ (i.e., in which one estimates $P(A)$ using $P_n(A)$ simultaneously for all $A \in \mathcal{A}$) is handled by the classical VC theory. Let us assume that both $\mathcal{A}, \mathcal{B}$ have VC dimension upper-bounded by some d_0.

To demonstrate the kinds of statements we would like, consider the case where each of $\mathcal{A}, \mathcal{B}$ contains only one event: $\mathcal{A} = \{A\}$, and $\mathcal{B} = \{B\}$, and set $\#_n(B) = \sum_i 1_{[x_i \in B]}$. A Chernoff bound implies that conditioned on the event that $\#_n(B) > 0$, the following holds with probability at least $1 - \delta$:

$$|P(A|B) - P_n(A|B)| \leq \sqrt{\frac{2\log(1/\delta)}{\#_n(B)}}. \tag{10}$$

This bound depends on $\#_n(B)$ and is thus data-dependent. To derive it, use that conditioned on $x_i \in B$, event $x_i \in A$ has probability $P(A|B)$, so random variable "$\#_n(B) \cdot p_n(A|B)$" has a binomial distribution with parameters $\#_n(B)$ and $P(A|B)$.

We would want to prove a uniform version of (10), of the form: with probability at least $1 - \delta$,

$$(\forall A \in \mathcal{A})\,(\forall B \in \mathcal{B}) : |P(A|B) - P_n(A|B)| \leq O\left(\sqrt{\frac{d_0 \log(1/\delta)}{\#_n(B)}}\right).$$

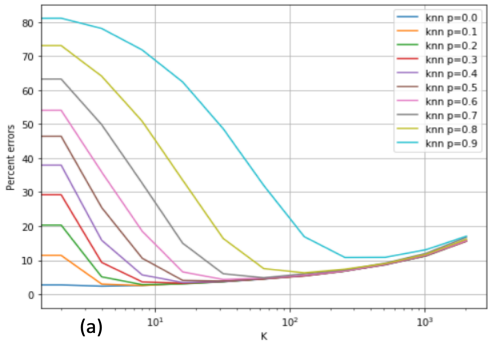
But as we explain in the appendix, this is unfortunately false. Instead, we prove the following (slightly weaker) variant:

Theorem 5 (UCECM)**.** *Let P be a probability distribution over X, and let $\mathcal{A}, \mathcal{B}$ be two families of measurable subsets of X such that $\mathsf{VC}(\mathcal{A}), \mathsf{VC}(\mathcal{B}) \leq d_0$. Let $n \in \mathbb{N}$, and let $x_1 \ldots x_n$ be n i.i.d samples from P. Then the following event occurs with probability at least $1 - \delta$:*

$$(\forall A \in \mathcal{A})\,(\forall B \in \mathcal{B}) : |P(A|B) - P_n(A|B)| \leq \sqrt{\frac{k_o}{\#_n(B)}},$$

where $k_o = 1000\,(d_0 \log(8n) + \log(4/\delta))$, and $\#_n(B) = \sum_{i=1}^n 1[x_i \in B]$.

7 Experiments



noise level	optimal k	error of optimal k	error of adaptive knn
0.0	4	2.39	3.39
0.1	8	2.59	3.61
0.2	8	2.86	3.79
0.3	16	3.27	4.26
0.4	16	3.46	4.53
0.5	32	3.91	5.12
0.6	32	4.35	6.15
0.7	64	4.87	7.71
0.8	128	6.32	9.50
0.9	256	10.84	12.22

(b)

Figure 3: Effect of label noise on k-NN and AKNN. Performance on MNIST for different levels of random label noise p and for different values of k. Each line in the figure on the left **(a)** represents the performance of k-NN as a function of k for a given level of noise. The optimal choice of k increases with the noise level, and that the performance degrades severely for too-small k. The table **(b)** shows that AKNN, with a fixed value of A, performs almost as well as k-NN with the optimal choice of k.

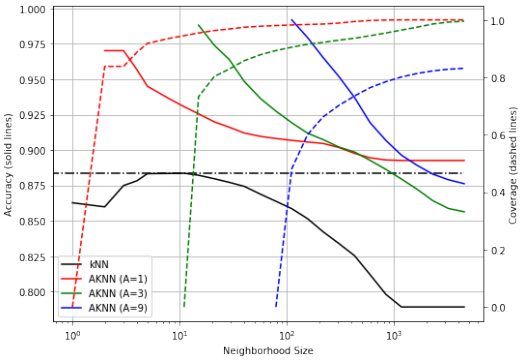
We performed a few experiments using real-world data sets from computer vision and genomics (see Section C). These were conducted with some practical alterations to the algorithm of Fig. 2.

Multiclass extension: Suppose the set of possible labels is $\mathcal{Y}$. We replace the binary rule "find the smallest k such that $|\eta_n(B_k(x))| > \Delta(n,k,\delta)$" with the rule: "find the smallest k such that $\eta_n^y(B_k(x)) - \frac{1}{|\mathcal{Y}|} > \Delta(n,k,\delta)$ for some $y \in \mathcal{Y}$, where $\eta_n^y(S) \doteq \frac{\#_n\{x_i \in S \text{ and } y_i = y\}}{\#_n(S)}$."

Parametrization: We replace Equation (6) with $\Delta = \frac{A}{\sqrt{k}}$, where A is a confidence parameter corresponding to the theory's δ (given n).

Resolving multilabel predictions: Our algorithm can output answers that are not a single label. The output can be "?", which indicates that no label has sufficient evidence. It can also be a subset of $\mathcal{Y}$ that contains more than one element, indicating that more than one label has significant evidence. In some situations, using subsets of the labels is more informative. However, when we want to compare head-to-head with k-NN, we need to output a single label. We use a heuristic to predict with a single label $y \in \mathcal{Y}$ on any x: the label for which $\max_k \eta_n^y(B_k(x))/\sqrt{k}$ is largest.

We briefly discuss our main conclusions from the experiments, with more details in Appendix C.



At left: performance of AKNN on notMNIST for different settings of the confidence parameter ($A = 1, 3, 9$), as a function of the neighborhood size. For each confidence level we show two graphs: an accuracy graph (solid line) and a coverage line (dashed line). For each value of k we plot the accuracy and the coverage of AKNN which is restricted to using a neighborhood size of at most k. Increasing A generally causes an increase in the accuracy and a decrease in coverage. Larger values of A cause AKNN to have coverage zero for values of k that are too small. For comparison, we plot the performance of k-NN as a function of k. The highest accuracy (≈ 0.88) is achieved for $k = 10$ (dotted horizontal line), and is surpassed by AKNN with high coverage (100% for $A = 1$).

Figure 4: Performance of AKNN on notMNIST. See also Figure 5.

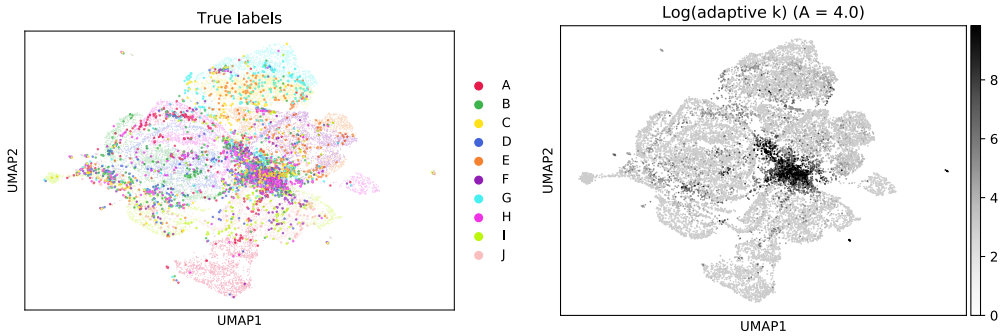


Figure 5: A visualization of the performance of AKNN on notMNIST. **(a)** The correct labels, with prediction errors of AKNN ($A = 4$) highlighted. **(b)** The value of k chosen by the algorithm when predicting each datapoint.

AKNN is comparable to the best k-NN rule. In Section 4.2 we prove that AKNN compares favorably to k-NN with any fixed k. We demonstrate this in practice in different situations. With simulated independent label noise on the MNIST dataset (Fig. 3), a small value of k is optimal for noiseless data, but performs very poorly when the noise level is high. On the other hand, AKNN adapts to the local noise level automatically, as demonstrated without adding noise on the more challenging notMNIST and single-cell genomics data (Fig. 4, 5, 6).

Varying the confidence parameter A controls abstaining. The parameter A controls how conservative the algorithm is in deciding to abstain, instead of incurring error by predicting. $A \to 0$ represents the most aggressive setting, in which the algorithm never abstains, essentially predicting according to a 1-NN rule. Higher settings of A cause the algorithm to abstain on some of these predicted points, for which there is no sufficiently small neighborhood with a sufficiently significant label bias (Fig. 7).

Adaptively chosen neighborhood sizes reflect local confidence. The number of neighbors chosen by AKNN is a local quantity that gives a practical pointwise measure of the confidence associated with label predictions. Small neighborhoods are chosen when one label is measured as significant nearly as soon as statistically possible; by definition of the AKNN stopping rule, this is not true where large neighborhoods are necessary. In our experiments, performance on points with significantly higher neighborhood sizes dropped monotonically, with the majority of the data set having performance significantly exceeding the best k-NN rule over a range of settings of A (Fig. 4, 6; Appendix C).

References

[AT07] J.-Y. Audibert and A.B. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35(2):608–633, 2007.

[BBL05] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[C+18] Tabula Muris Consortium et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367, 2018.

[CD10] K. Chaudhuri and S. Dasgupta. Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems*, pages 343–351, 2010.

[CD14] K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445. 2014.

[CG06] F. Cerou and A. Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

[CH67] T. Cover and P.E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967.

[CS18] G.H. Chen and D. Shah. *Explaining the Success of Nearest Neighbor Methods in Prediction*. Foundations and Trends in Machine Learning. NOW Publishers, 2018.

[DCL11] W. Dong, M. Charikar, and K. Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586. ACM, 2011.

[DGKL94] L. Devroye, L. Györfi, A. Krzyzak, and G. Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *Annals of Statistics*, 22:1371–1385, 1994.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[Dud79] R.M. Dudley. Balls in $\mathbb{R}^k$ do not cut all subsets of $k+2$ points. *Advances in Mathematics*, 31(3):306–308, 1979.

[FH51] E. Fix and J. Hodges. Discriminatory analysis, nonparametric discrimination. *USAF School of Aviation Medicine, Randolph Field, Texas, Project 21-49-004, Report 4, Contract AD41(128)-31*, 1951.

[Gyö81] L. Györfi. The rate of convergence of k_n-nn regression estimates and classification rules. *IEEE Transactions on Information Theory*, 27(3):362–364, 1981.

[Hei01] J. Heinonen. *Lectures on Analysis on Metric Spaces*. Springer, 2001.

[KP95] S. Kulkarni and S. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

[Kpo11] S. Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Neural Information Processing Systems*, 2011.

[MNI96] MNIST dataset. http://yann.lecun.com/exdb/mnist/, 1996.

[Mou18] Mouse cell atlas dataset. `ftp://ngs.sanger.ac.uk/production/teichmann/BBKNN/MouseAtlas.zip`, 2018. Accessed: 2019-05-02.

[MT99] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[not11] notMNIST dataset. `http://yaroslavb.com/upload/notMNIST/`, 2011. Accessed: 2019-05-02.

[RS98] M. Raab and A. Steger. Balls into bins - a simple and tight analysis. In *Randomization and Approximation Techniques in Computer Science, Second International Workshop, RANDOM'98, Barcelona, Spain, October 8-10, 1998, Proceedings*, pages 159–170, 1998.

[Sto77] C. Stone. Consistent nonparametric regression. *Annals of Statistics*, 5:595–645, 1977.

[VC71] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

A Analysis and proofs

The first step in establishing advantage-dependent rates of convergence is to bound the accuracy of empirical estimates of probability mass and bias. This is achieved by a careful choice of large deviation bounds.

A.1 Large deviation bounds

Suppose we draw n points $(x_1, y_1), \ldots, (x_n, y_n)$ from P. If n is reasonably large, we would expect the empirical mass $\mu_n(S)$ of any set $S \subset \mathcal{X}$, as defined in (4), to be close to its probability mass under μ. The following lemma, from [CD10], quantifies one particular aspect of this.

Lemma 6 ([CD10], Lemma 7)**.** *There is a universal constant c_0 such that the following holds. Let $\mathcal{B}$ be any class of measurable subsets of $\mathcal{X}$ of VC dimension d_0. Pick any $0 < \delta < 1$. Then with probability at least $1 - \delta^2/2$ over the choice of $(x_1, y_1), \ldots, (x_n, y_n)$, for all $B \in \mathcal{B}$ and for any integer k, we have*

$$\mu(B) \geq \frac{k}{n} + \frac{c_0}{n} \max\left(k, d_0 \log \frac{n}{\delta}\right) \quad \Longrightarrow \quad \mu_n(B) \geq \frac{k}{n}.$$

Likewise, we would expect the empirical bias $\eta_n(S)$ of a set $S \subset \mathcal{X}$, as defined in (5), to be close to its true bias $\eta(S)$. The latter is defined whenever $\mu(S) > 0$.

Lemma 7. *There is a universal constant c_1 for which the following holds. Let $\mathcal{C}$ be a class of subsets of $\mathcal{X}$ with VC dimension d_0. Pick any $0 < \delta < 1$. Then with probability at least $1 - \delta^2/2$ over the choice of $(x_1, y_1), \ldots, (x_n, y_n)$, for all $C \in \mathcal{C}$,*

$$|\eta_n(C) - \eta(C)| \ \leq \ \Delta(n, \#_n(C), \delta)$$

where $\#_n(C) = |\{i : x_i \in B\}|$ is the number of points in C and

$$\Delta(n, k, \delta) = c_1 \sqrt{\frac{d_0 \log n + \log(1/\delta)}{k}}. \tag{11}$$

Lemma 7 is a special case[1] of a uniform convergence bound for conditional probabilities (Theorem 5) that we prove in Section 6.

A.2 Proof of Theorem 1

Theorem 1 is an immediate consequence of the following lemma, in which the choice of constants is made explicit.

Lemma 8. *Define $c_2 = \max(c_1, 1/2)\sqrt{1 + c_0}$, where c_0 and c_1 are the constants from Lemmas 6 and 7, and take $c_3 = 16c_2^2$. Pick any $x \in \text{supp}(\mu)$ with $\text{adv}(x) > 0$. Fix any $0 < \delta < 1$. If the number of training points satisfies*

$$n > \frac{c_3}{\text{adv}(x)} \max\left(\log \frac{c_3}{\text{adv}(x)}, \ \log \frac{1}{\delta}\right),$$

then with probability at least $1 - \delta^2$ over the choice of training data, the adaptive nearest neighbor rule will have $g_n(x) = g^*(x)$.

Proof. Pick any $x \in \text{supp}(\mu)$. Suppose $\eta(x) > 0$; the negative case is symmetric. The set $\mathcal{B}$ of all balls centered at x is easily seen to have VC dimension $d_0 = 1$. By Lemmas 6 and 7, we have that with probability at least $1 - \delta^2$, the following two properties hold for all $B \in \mathcal{B}$:

1. For any integer k, we have $\#_n(B) \geq k$ whenever $n\mu(B) \geq k + c_0 \max(k, \log(n/\delta))$.
2. $|\eta_n(B) - \eta(B)| \leq \Delta(n, \#_n(B), \delta)$.

[1] Indeed, Lemma 7 follows from Theorem 5 by plugging in $\mathcal{A} = \{\mathcal{X} \times \{+1\}\}, \mathcal{B} = \{C \times \{\pm 1\} : C \in \mathcal{C}\}$.

Assume henceforth that these hold.

By the definition of advantage, point x is (p, γ)-salient for some $p, \gamma > 0$ with $p\gamma^2 = \mathrm{adv}(x) - \epsilon$, where we can make $\epsilon > 0$ arbitrarily small. The lower bound on n in the theorem statement then implies that

$$\gamma \geq 2c_2 \sqrt{\frac{\log n + \log(1/\delta)}{np}}, \tag{12}$$

or equivalently that $np\gamma^2 \geq 4c_2^2(\log n + \log(1/\delta))$.

Set $k = np/(1 + c_0)$. By (12) we have $np \geq 4c_2^2 \log(n/\delta)$ and thus $k \geq \log(n/\delta)$. As a result, $np \geq k + c_0 \max(k, \log(n/\delta))$, and by property 1, the ball $B = B(x, r_p(x))$ has $\#_n(B) \geq k$. This means, in turn, that by property 2,

$$\begin{aligned}
\eta_n(B) \geq\ & \eta(B) - \Delta(n, k, \delta) = \gamma - c_1 \sqrt{\frac{\log(n/\delta)}{k}} \\
& \geq 2c_2 \sqrt{\frac{\log(n/\delta)}{np}} - c_1 \sqrt{\frac{\log(n/\delta)}{k}} \geq 2c_1 \sqrt{\frac{\log(n/\delta)}{k}} - c_1 \sqrt{\frac{\log(n/\delta)}{k}} \\
& = c_1 \sqrt{\frac{\log(n/\delta)}{k}} \geq \Delta(n, \#_n(B), \delta).
\end{aligned}$$

Thus ball B would trigger a prediction of $+1$.

At the same time, for any ball $B' = B(x, r)$ with $r < r_p(x)$,

$$\eta_n(B') \geq \eta(B') - \Delta(n, \#_n(B'), \delta) > -\Delta(n, \#_n(B'), \delta)$$

and thus no such ball will trigger a prediction of -1. Therefore, the prediction at x must be $+1$. $\square$

A.3 Proof of Theorem 2

This proof follows much the same outline as that of Theorem 1. A crucial difference is that uniform large deviation bounds (Lemmas 6 and 7) are applied to the class of all balls in $\mathcal{X}$, which is assumed[2] to have finite VC dimension d_0. In contrast, the proof of Theorem 1 only applies these bounds to the class of balls centered at a specific point, which has VC dimension at most 1 in any metric space.

A.4 Proof of Theorem 4

Recall from (8) that $\mathcal{X}_a$ denotes the set of points with advantage $> a$.

Lemma 9. *Let c_3 be the constant from Lemma 8. Pick any $0 < \delta < 1$ as a confidence parameter for the* AKNN *estimator of Figure 2. Fix any $a > 0$. If the number of training points n satisfies*

$$n \geq \frac{c_3}{a} \max\left(\log \frac{c_3}{a},\ \log \frac{1}{\delta}\right),$$

then with probability at least $1 - \delta$, the resulting classifier g_n has risk

$$R(g_n) - R^* \leq \delta + \mu(\mathcal{X}_0 \setminus \mathcal{X}_a).$$

Proof. From Lemma 8, we have that for any $x \in \mathcal{X}_a$,

$$\Pr_n(g_n(x) \neq g^*(x)) \leq \delta^2,$$

where $\Pr_n$ denotes probability over the choice of training points. Thus, for $X \sim \mu$,

$$\mathbb{E}_n \mathbb{E}_X 1(g_n(X) \neq g^*(X) | X \in \mathcal{X}_a) \leq \delta^2,$$

and by Markov's inequality,

$$\Pr_n[\Pr_X(g_n(X) \neq g^*(X) | X \in \mathcal{X}_a) \geq \delta] \leq \delta.$$

[2]This is motivated by finite-dimensional Euclidean space $\mathbb{R}^D$, where it holds with $d_0 = D + 1$ ([Dud79]).

Thus, with probability at least $1-\delta$ over the training set,

$$\Pr_X(g_n(X) \neq g^*(X) | X \in \mathcal{X}_a) \leq \delta.$$

On points with $\eta(x) = 0$, both g_n and the Bayes-optimal g^* incur the same risk. Thus

$$\begin{aligned} R(g_n) - R^* &\leq \Pr_X(g_n(X) \neq g^*(X) | X \in \mathcal{X}_a) + \Pr_X(X \notin \mathcal{X}_a, \eta(X) \neq 0) \\ &\leq \delta + \Pr_X(X \in \mathcal{X}_0 \setminus \mathcal{X}_a) + \Pr_X(\mathrm{adv}(X) = 0, \eta(X) \neq 0) \\ &\leq \delta + \mu(\mathcal{X}_0 \setminus \mathcal{X}_a), \end{aligned}$$

where we invoke Lemma 3 for the last step. □

We now complete the proof of Theorem 4. Given the sequence of confidence parameters (δ_n), define a sequence of advantage values (a_n) by

$$a_n = \frac{c_3}{n} \max\left(2 \log n,\ \log \frac{1}{\delta_n}\right).$$

The conditions on (δ_n) imply $a_n \to 0$.

Pick any $\epsilon > 0$. By the conditions on (δ_n), we can pick N so that $\sum_{n \geq N} \delta_n \leq \epsilon$. Let ω denote a realization of an infinite training sequence $(X_1, Y_1), (X_2, Y_2), \ldots$ from P. By Lemma 9, for any positive integer N,

$$\Pr\left(\omega : \exists n \geq N \text{ s.t. } R(g_n(\omega)) - R^* > \delta_n + \mu(\mathcal{X}_0 \setminus \mathcal{X}_{a_n})\right) \leq \sum_{n \geq N} \delta_n \leq \epsilon.$$

Thus, with probability at least $1-\epsilon$ over the training sequence ω, we have that for all $n \geq N$,

$$R(g_n(\omega)) - R^* \leq \delta_n + \mu(\mathcal{X}_0 \setminus \mathcal{X}_{a_n}),$$

whereupon $R(g_n(\omega)) \to R^*$ (since $\delta_n, a_n \to 0$ and $\lim_{a \downarrow 0} \mu(\mathcal{X}_0 \setminus \mathcal{X}_a) = 0$). Since this holds for any $\epsilon > 0$, the theorem follows.

B Uniform Convergence of Empirical Conditional Measures

B.1 Formal Statement

Let P be a distribution over X, and let $\mathcal{A}, \mathcal{B}$ be two collections of events. Consider n independent samples from P, denoted by $x_1, \ldots, x_n$. We would like to estimate $P(A|B)$ simultaneously for all $A \in \mathcal{A}, B \in \mathcal{B}$. It is natural to consider the empirical estimates:

$$P_n(A|B) = \frac{\sum_i 1_{[x_i \in A \cap B]}}{\sum_i 1_{[x_i \in B]}}.$$

We study when (and to what extent) these estimates provide a good approximation. Note that the case where $\mathcal{B} = \{X\}$ (i.e., in which one estimates $P(A)$ using $P_n(A)$ simultaneously for all $A \in \mathcal{A}$) is handled by the classical VC theory. Throughout this section we assume that both $\mathcal{A}, \mathcal{B}$ have finite VC dimension, and we let d_0 denote an upper bound on both $\mathsf{VC}(\mathcal{A})$ and $\mathsf{VC}(\mathcal{B})$.

To demonstrate the kinds of statements we would like to derive, consider the case where each of $\mathcal{A}, \mathcal{B}$ contains only one event: $\mathcal{A} = \{A\}$, and $\mathcal{B} = \{B\}$, and set $\#_n(B) = \sum_i 1_{[x_i \in B]}$. A Chernoff bound implies that conditioned on the event that $\#_n(B) > 0$, the following holds with probability at least $1-\delta$:

$$|P(A|B) - P_n(A|B)| \leq \sqrt{\frac{2 \log(1/\delta)}{\#_n(B)}}. \tag{13}$$

To derive it, use that conditioned on $x_i \in B$, the event $x_i \in A$ has probability $P(A|B)$, and therefore the random variable "$\#_n(B) \cdot p_n(A|B)$" has a binomial distribution with parameters $\#_n(B)$ and $P(A|B)$.

Note that the bound on the error in Equation (13) depends on $\#_n(B)$ and therefore is data-dependent. We stress that this is the type of statement we want: the more samples belong to B, the more certain we are with the empirical estimate. Thus, we would want to prove a statement as follows:

With probability at least $1-\delta$,

$$(\forall A \in \mathcal{A})\,(\forall B \in \mathcal{B}) : |P(A|B) - P_n(A|B)| \leq O\left(\sqrt{\frac{d_0 \log(1/\delta)}{\#_n(B)}}\right),$$

where $\#_n(B) = \sum_{i=1}^{n} 1[x_i \in B]$.

The above statement is, unfortunately, false. As an example, consider the probability space defined by drawing $x \sim [n]$ uniformly, and then coloring x by $c_x \in \{\pm 1\}$ uniformly. For each i let B_i denote the event that i was drawn, and let A denote the event that the drawn color was $+1$. (formally, $B_i = \{i\} \times \{\pm 1\}$, and $A = [n] \times \{+1\}$). One can verify that the VC dimension of $\mathcal{B} = \{B_i : i \leq n\}$ and of $\mathcal{A} = \{A\}$ is at most 1. The above statement fails in this setting: indeed, one can verify that if we draw n samples from this space then with a constant probability there will be some j such that:

(i) j always gets the same color (say $+1$), and

(ii) j is sampled at least $\Omega(\log n / \log\log n)$ times[3].

Therefore, with constant probability we get that

$$P_n(A|B_i) = 1, P(A|B_i) = 1/2,$$

and so the difference between the error is clearly $1-(1/2) = 1/2$, which is clearly not upper bounded by $O(\sqrt{\log\log n / \log n})$.

We prove the following (slightly weaker) variant:

Theorem (Theorem 5 restatement)**.** *Let P be a probability distribution over X, and let $\mathcal{A}, \mathcal{B}$ be two families of measurable subsets of X such that $\mathsf{VC}(\mathcal{A}), \mathsf{VC}(\mathcal{B}) \leq d_0$. Let $n \in \mathbb{N}$, and let $x_1 \ldots x_n$ be n i.i.d samples from P. Then the following event occurs with probability at least $1-\delta$:*

$$(\forall A \in \mathcal{A})\,(\forall B \in \mathcal{B}) : |P(A|B) - P_n(A|B)| \leq \sqrt{\frac{k_o}{\#_n(B)}},$$

where $k_o = 1000\,(d_0 \log(8n) + \log(4/\delta))$, and[4] *$\#_n(B) = \sum_{i=1}^{n} 1[x_i \in B]$.*

Discussion. Theorem 5 can be combined with Lemma 6 to yield a bound on the minimal n for which $P_n(A|B)$ is a non-trivial approximation of $P(A|B)$. Indeed, Lemma 6 implies that if n is large enough so that $P(B) = \Omega\left(\frac{d_0 \log n}{n}\right)$, then the empirical estimate $P_n(A|B)$ is a decent approximation. In the context of the adaptive nearest neighbor classifier, this means that the empirical biases provide meaningful estimates of the true biases for balls whose measure is $\tilde{\Omega}\left(\frac{d_0}{n}\right)$. This resembles the learning rate in realizable settings.

We remark that a weaker statement than Theorem 5 can be derived as a corollary of the classical uniform convergence result [VC71]. Indeed, since the VC dimension of $\{B \cap A : i \in \mathcal{I}\}$ is at most d_0, it follows that

$$P_n(A|B) \approx \frac{P(A \cap B) \pm \sqrt{d_0/n}}{P(B) \pm \sqrt{d_0/n}}.$$

However, this bound guarantees non-trivial estimates only once $P(B)$ is roughly $\sqrt{d_0/n}$. This is similar to the learning rate in agnostic (i.e., non-realizable) settings.

Another major advantage of the uniform convergence bound in Theorem 5 is that it is data-dependent: if many points from the sample belong to $B \in \mathcal{B}$ (i.e. $\#_n(B)$ is large), then we get better guarantees on the approximation of $P(A|B)$ by $P_n(A|B)$ for all $A \in \mathcal{A}$.

[3] This follows from analyzing the maximal bin in a uniform assignment of $\Theta(n)$ balls into n bins [RS98]

[4] Note that the above inequality makes sense also when $k(B) = 0$, by identifying $\frac{\cdot}{0}$ as ∞, and using the convention that $\infty - \infty = \infty$ and that $\infty \leq \infty$.

B.2 Proof of Theorem 5

As noted above, the standard uniform convergence bound for VC classes can not yield the bound in Theorem 5. Instead, we use a variant of it due to [BBL05] which concerns *relative deviations* (see [BBL05]: Theorem 5.1 and the discussion before Corollary 5.2). In order to state the theorem, we need the following notation: Let $\mathcal{C}$ be a family of subsets of $\mathcal{X}$. We denote by $\mathbb{S}_{\mathcal{C}} : \mathbb{N} \to \mathbb{N}$ the *growth function* of $\mathcal{C}$, which is defined by:

$$\mathbb{S}_{\mathcal{C}}(n) = \max\{|\mathcal{C}|_R| : R \subseteq X, |R| = n\},$$

where $\mathcal{C}|_R = \{C \cap R : C \in \mathcal{C}\}$ is the projection of $\mathcal{C}$ to R.

Theorem 10 ([BBL05])**.** *Let $\mathcal{C}$ be a family of subsets of $\mathcal{X}$ and let P be a distribution over $\mathcal{X}$. Then, the following holds with probability $1-\delta$:*

$$(\forall C \in \mathcal{C}) : |P(C) - P_n(C)| \leq 2\sqrt{P_n(C)\frac{\log \mathbb{S}_{\mathcal{C}}(2n) + \log(4/\delta)}{n}} + 4\frac{\log \mathbb{S}_{\mathcal{C}}(2n) + \log(4/\delta)}{n}.$$

Set $\mathcal{C} = \mathcal{B} \cup \{A \cap B : A \in \mathcal{A}, B \in \mathcal{B}\}$. We prove Theorem 5 by applying Theorem 10 on $\mathcal{C}$; to this end we first upper bound $\mathbb{S}_{\mathcal{C}}(n)$. Let $\mathcal{D} = \{A \cap B : A \in \mathcal{A}, B \in \mathcal{B}\}$, so that $\mathcal{C} = \mathcal{B} \cup \mathcal{D}$. Then:

$$\mathbb{S}_{\mathcal{C}}(n) \leq \mathbb{S}_{\mathcal{B}}(n) + \mathbb{S}_{\mathcal{D}}(n) \leq \mathbb{S}_{\mathcal{B}}(n) + \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n) \leq 2\mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n) \leq 2\binom{n}{\leq d_0}^2 \leq 2(2n)^{2d_0},$$

where the second inequality follows since $\mathbb{S}_{\mathcal{D}}(n) \leq \mathbb{S}_{\mathcal{A}}(n)\mathbb{S}_{\mathcal{B}}(n)$, the second to last inequality follows from the Sauer-Shelah-Perles Lemma, and the last inequality follows since $\binom{a}{\leq b} \leq (2a)^b$. Therefore, applying Theorem 10 on $\mathcal{C}$ yields that with probability $1-\delta$ the following event holds:

$$(\forall C \in \mathcal{C}) : |P(C) - P_n(C)| \leq 4\sqrt{P_n(C)\frac{d_0 \log 8n + \log(4/\delta)}{n}} + 8\frac{d_0 \log 8n + \log(4/\delta)}{n}. \quad (14)$$

For the remainder of the proof we assume that the event in Equation (14) holds and argue that it implies the conclusion in Theorem 5. Let $A \in \mathcal{A}, B \in \mathcal{B}$, and let $k = n \cdot P_n(B) = \#_n(B)$ denote the number of data points in B. We want to show that

$$|P(A|B) - P_n(A|B)| \leq \sqrt{\frac{k_o}{k}}, \quad (15)$$

where $k_o = 1000\,(d_0 \log(8n) + \log(4/\delta))$. Let $j = k \cdot P_n(A|B) = \#_n(A \cap B)$ denote the number of data points in $A \cap B$. We establish Equation (15) by showing that

$$P(A|B) \leq P_n(A|B) + \sqrt{\frac{k_o}{k}} \quad \text{and} \quad P(A|B) \geq P_n(A|B) - \sqrt{\frac{k_o}{k}}.$$

In the following calculation it will be convenient to denote $D := d_0 \log(8n) + \log(4/\delta)$. By Equation (14) we get:

$$\begin{aligned}
P(A|B) &= \frac{P(A \cap B)}{P(B)} \\
&\leq \frac{P_n(A \cap B) + 4\sqrt{P_n(A \cap B)\frac{D}{n}} + 8\frac{D}{n}}{P_n(B) - 4\sqrt{P_n(B)\frac{D}{n}} - 8\frac{D}{n}} \\
&= \frac{\frac{P_n(A\cap B)}{P_n(B)} + 4\sqrt{\frac{P_n(A\cap B)}{P_n(B)}\frac{D}{nP_n(B)}} + 8\frac{D}{nP_n(B)}}{1 - 4\sqrt{\frac{D}{nP_n(B)}} - 8\frac{D}{nP_n(B)}} s = P_n(A|B)\frac{1 + 4\sqrt{\frac{D}{j}} + 8\frac{D}{j}}{1 - 4\sqrt{\frac{D}{k}} - 8\frac{D}{k}},
\end{aligned}$$

where the first inequality follows from Equation (14) and the following equalities are trivial. Thus,

$$P(A|B) \leq \frac{j}{k}\left(\frac{1 + 4\sqrt{\frac{D}{j}} + 8\frac{D}{j}}{1 - 4\sqrt{\frac{D}{k}} - 8\frac{D}{k}}\right). \quad (16)$$

Next, note that we may assume that $k \geq k_o = 1000D$, as otherwise Equation (15) trivially holds. Therefore,

$$\frac{1}{1-4\sqrt{\frac{D}{k}}-8\frac{D}{k}} \leq 1+8\sqrt{\frac{D}{k}}+16\frac{D}{k}. \qquad ((\forall x < \tfrac{1}{2}) : \tfrac{1}{1-x} \leq 1+2x)$$

Plugging this in Equation (16), and using first that $j \leq k$ and then that $1000D \leq k$, yields:

$$\begin{aligned}
P(A|B) &\leq \frac{j}{k}\Big(1+4\sqrt{\frac{D}{j}}+8\frac{D}{j}\Big)\Big(1+8\sqrt{\frac{D}{k}}+16\frac{D}{k}\Big) \\
&= \frac{j}{k}+8\frac{j}{k}\sqrt{\frac{D}{k}}\left(1+2\sqrt{\frac{D}{k}}\right)+\Big(\frac{4\sqrt{jD}+8D}{k}\Big)\Big(1+4\sqrt{\frac{D}{k}}\Big)^2 \\
&\leq \frac{j}{k}+8\sqrt{\frac{D}{k}}\left(1+2\sqrt{\frac{D}{k}}\right)+\Big(4\sqrt{\frac{D}{k}}+\frac{8D}{k}\Big)\Big(1+4\sqrt{\frac{D}{k}}\Big)^2 \\
&\leq \frac{j}{k}+30\sqrt{\frac{D}{k}} = P_n(A|B)+\sqrt{\frac{k_o}{k}},
\end{aligned}$$

and so

$$P(A|B) \leq P_n(A|B)+\sqrt{\frac{k_o}{k}}.$$

A symmetric argument yields similarly to Equation (16) that:

$$P(A|B) \geq \frac{j}{k}\left(\frac{1-4\sqrt{\frac{D}{j}}-8\frac{D}{j}}{1+4\sqrt{\frac{D}{k}}+8\frac{D}{k}}\right).$$

Then, a similar calculation (using the relation $(\forall x > 0) : \frac{1}{1+x} \geq 1-2x$) implies that

$$P(A|B) \geq P_n(A|B)-\sqrt{\frac{k_o}{k}},$$

which finishes the proof. □

C Experimental Results

C.1 Datasets

We test AKNN on several datasets. The first was the MNIST dataset of 70000 examples ([MNI96]).

We also evaluate AKNN on the more challenging notMNIST dataset ([not11]), consisting of extracted glyphs of the letters A-J from publicly available fonts. We use the 18724 labeled examples from this set, preprocessed feature-wise to be in $[-\frac{1}{2}, \frac{1}{2}]$ using $x \mapsto \frac{x}{255} - \frac{1}{2}$.

We further use AKNN on a challenging binary classification task of continuing interest, involving gene expression data on a population of single cells from different mouse organs collected by the Tabula Muris consortium ([C$^+$18], as processed in [Mou18]). This constitutes 45291 cells (training examples). Each cell has its data collected using one of two approaches. The task is to classify between them. More details follow.

The data are collected using representative protocols of two currently dominant approaches to isolate and measure single cells: a plate-based approach sorting cells into microwells, and a droplet-based approach manipulating cells using microfluidic technologies. Each approach has its own set of technical biases, about which much remains to be understood. Identifying and characterizing these biases to discriminate between such approaches is currently of great interest.

Both approaches measure effectively the same cells for our purposes, so there is a large decision boundary in the binary classification problem.

C.2 A note on efficient implementation

In this paper, we computed the nearest neighbors of data exactly when running AKNN, to faithfully demonstrate its behavior. In practice, this would be done using approximate nearest-neighbor search to build a k-NN graph using a small fixed k (say 10), and then using pairwise distances on this graph to compute neighborhoods as needed by AKNN. We tried this (using the nearest-neighbor method of [DCL11]) on notMNIST without substantive differences in the results.

C.3 Supplemental Figures

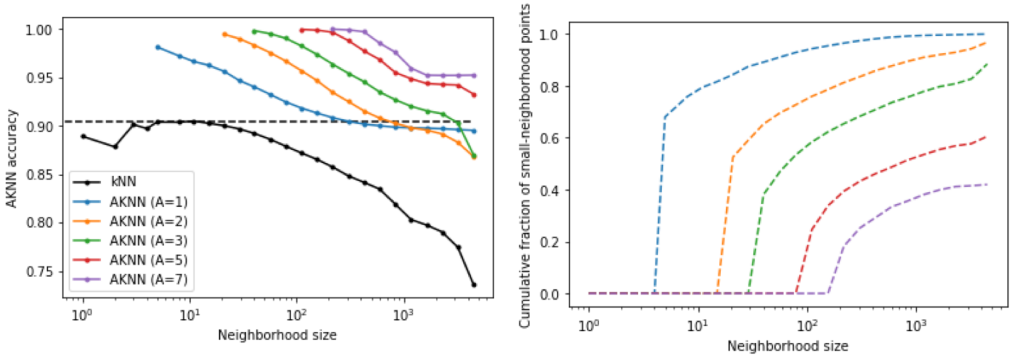


Figure 6: As Fig. 4, on single-cell mouse data. AKNN is notably accurate on small-neighborhood points at moderate coverage, and performance drops off at higher k, with A controlling this frontier.

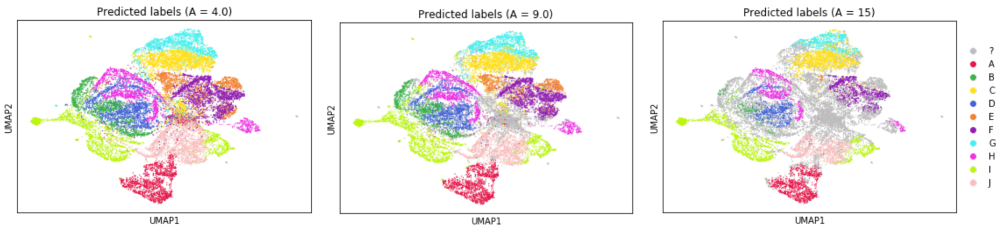


Figure 7: AKNN predictions on notMNIST, for different settings of A.

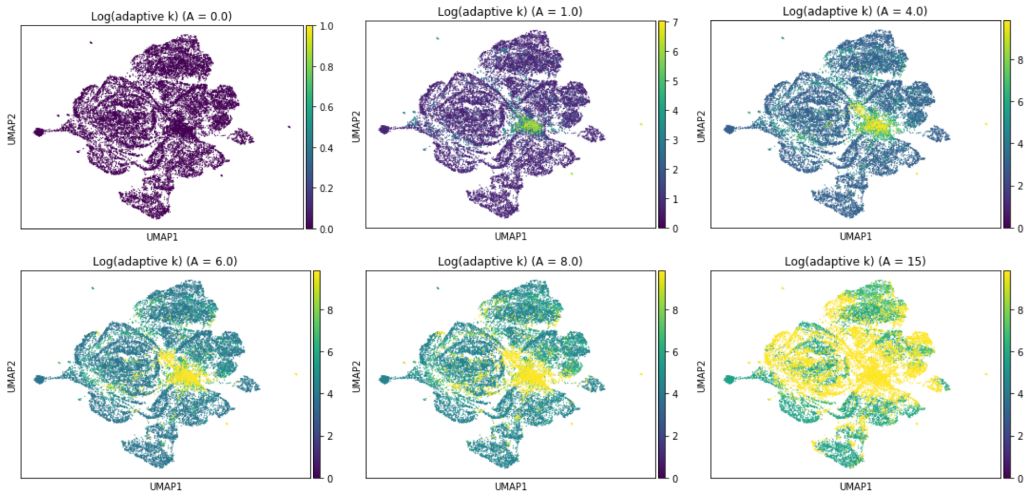


Figure 8: AKNN neighborhood sizes on notMNIST, in increasing order of A, plotted on a log scale. Top left figure ($A = 0$) represents a 1-NN classifier. Bottom right figure ($A = 15$) shows that many of the points' neighborhoods are maximally large, which can be compared to the right panel of Fig. 7.

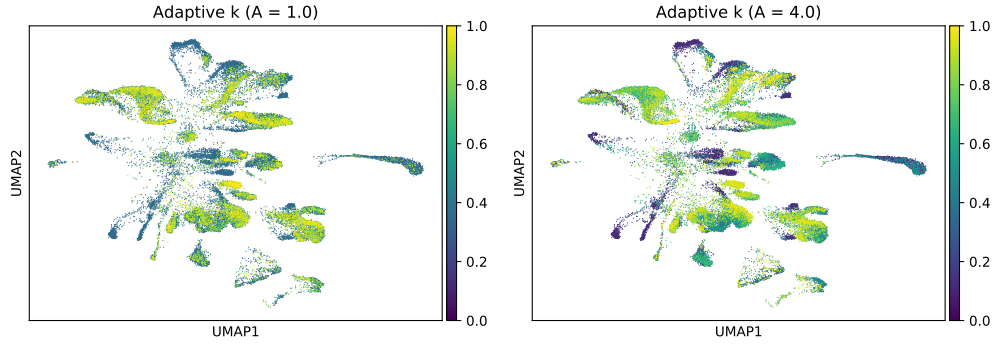


Figure 9: As Fig. 8, on single-cell mouse data, with the AKNN k-values replaced by their quantiles over the data. The relative ordering of the data by AKNN neighborhood size is fairly robust to A.