Achieving the time of 1-NN, but the accuracy of k-NN

Lirong Xue
Princeton University

Samory Kpotufe
Princeton University

Abstract

We propose a simple approach which, given distributed computing resources, can nearly achieve the accuracy of k-NN prediction, while matching (or improving) the faster prediction time of 1-NN. The approach consists of aggregating *denoised* 1-NN predictors over *a small number* of distributed subsamples. We show, both theoretically and experimentally, that small subsample sizes suffice to attain similar performance as k-NN, without sacrificing the computational efficiency of 1-NN.

1 INTRODUCTION

While k-Nearest Neighbor (k-NN) classification or regression can achieve significantly better prediction accuracy than 1-NN ($k = 1$), practitioners often default to 1-NN as it can achieve much faster prediction that scales better with large sample size n. In fact, much of the commercial tools for nearest neighbor search remain optimized for 1-NN rather than for k-NN, further biasing practice towards 1-NN. Unfortunately, 1-NN is statistically inconsistent, i.e., its prediction accuracy plateaus early as sample size n increases, while k-NN keeps improving longer for choices of $k \xrightarrow{n\to\infty} \infty$.

In this work we consider having access to *a small number* of distributed computing units, and ask whether better tradeoffs between k-NN and 1-NN can be achieved by harnessing parallelism at prediction time. A simple idea is *bagging* multiple 1-NN predictors computed over distributed subsamples; however this tends to require a large number of subsamples, while the number of computing units is often constrained in practice. In fact, an *infinite* number of subsamples is assumed in all known consistency guarantees for the 1-NN bagging approach (Biau et al., 2010; Samworth et al., 2012).

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

Here, we are particularly interested in small numbers of distributed subsamples (say 1 to 10) as a practical matter. Hence, we consider a simple variant of the above idea, consisting of aggregating a few *denoised* 1-NN predictors. With this simple change, we obtain the same theoretical error-rate guarantees as for k-NN, using fewer subsamples, while individual processing times are of the same order or better than 1-NN's computation time.

The main intuition behind denoising is as follows. The increase in variance due to subsampling is hard to counter if too few predictors are aggregated. We show that this problem is suitably addressed by *denoising* each subsample as a preprocessing step, i.e., replacing the subsample labels with k-NN estimates based on the original data. Prediction then consists of aggregating – by averaging or by majority voting – the 1-NN predictions from a few denoised subsamples (of small size $m \ll n$).

Interestingly, as shown both theoretically and experimentally, we can let the subsampling ratio $(m/n) \xrightarrow{n\to\infty} 0$ while achieving a prediction accuracy of the same order as that of k-NN. Such improved accuracy over vanilla 1-NN is verified experimentally, even for relatively small number of distributed predictors. Note that, in practice, we aim to minimize the number of distributed predictors, or equivalently the number of computing units which is usually costly in its own right. This is therefore a main focus in our experiments. In particular, we will see that even with a single denoised 1-NN predictor, i.e., one computer, we can observe a significant improvement in accuracy over vanilla 1-NN while maintaining the prediction speed of 1-NN. Our main focus in this work is classification – perhaps the most common form of NN prediction – but our results readily extend to regression.

Detailed Results And Related Work

While nearest neighbor prediction methods are among the oldest and most enduring in data analysis (Fix and Hodges Jr, 1951; Cover and Hart, 1967; Kulkarni and Posner, 1995), their theoretical performance in practical settings is still being elucidated. For statistical

consistency, it is well known that one needs a number $k \xrightarrow{n\to\infty} \infty$ of neighbors, i.e., the vanilla 1-NN method is inconsistent for either regression or classification (Devroye et al., 1994). In the case of regression, Kpotufe (2011) shows that convergence rates (l_2 excess error over Bayes) behave as $O(n^{-2/(2+d)})$, for Lipschitz regression functions over data with intrinsic dimension d; this then implies a rate of $O(n^{-1/(2+d)})$ for binary classification via known relations between regression and classification rates (see e.g. Devroye et al. (1996)). Similar rates are recovered in (Cannings et al., 2017) under much refined parametrization of the marginal input distribution, while a recent paper of Moscovich et al. (2016) recovers similar rates in semisupervised settings.

Such classification rates can be sharpened by taking into account the *noise margin*, i.e., the mass of data away from the decision boundary. This is done in the recent work of Chaudhuri and Dasgupta (2014) which obtain faster rates of the form $O(n^{-\alpha(\beta+1)/(2\alpha+d)})$ – where the regression function is assumed α-smooth – which can be much faster for large β (characterizing the noise margin). However such rates require large number of neighbors $k = O(n^{2\alpha/(2\alpha+d)})$ growing as a root of sample size n; such large k implies much slower prediction time in practice, which is exacerbated by the scarcity of optimized tools for 'k' nearest neighbor search. In contrast, fast commercial tools for 1-NN search are readily available, building on various space partitioning data structures (Krauthgamer and Lee, 2004; Clarkson, 2005; Beygelzimer et al., 2006; Gionis et al., 1999).

In this work we show that the classification error of the proposed approach, namely aggregated denoised 1-NN's, is of the same optimal order $\tilde{O}(n^{-\alpha(\beta+1)/(2\alpha+d)})$ plus a term $\tilde{O}(m^{-\alpha(\beta+1)/d})$ where $m \leq n$ is the subsample size used for each denoised 1-NN. This additional term, due to subsampling, is of lower order provided $m = \tilde{\Omega}(n^{d/(2\alpha+d)})$; in other words we can let the subsampling ratio $(m/n) = \tilde{\Omega}(n^{-2\alpha/(2\alpha+d)}) \xrightarrow{n\to\infty} 0$ while achieving the same rate as k-NN. We emphasize that the smaller the subsampling ratio, the faster the prediction time: rather than just maintaining the prediction time of vanilla 1-NN, we can actually get considerably better prediction time using smaller subsamples, while at the same time considerably improving prediction accuracy towards that of k-NN. Finally notice that the theoretical subsampling ratio of $\tilde{\Omega}(n^{-2\alpha/(2\alpha+d)})$ is best with smaller d, the *intrinsic* dimension of the data, which is not assumed to be known a priori. Such intrinsic dimension d is smallest for structured data in $\mathbb{R}^D$, e.g. data on an unknown manifold, or sparse data, and therefore suggests that much smaller subsamples – hence faster prediction times – are possible with structured data while achieving good prediction accuracy.

As mentioned earlier, the even simpler approach of *bagging* 1-NN predictors is known to be consistent (Biau and Devroye, 2010; Biau et al., 2010; Samworth et al., 2012), however only in the case of an infinite bag size, corresponding to an infinite number of computing units in our setting – we assume one subsample per computing unit so as to maintain or beat the prediction time of 1-NN. Interestingly, as first shown in (Biau and Devroye, 2010; Biau et al., 2010), the subsampling ratio (m/n) can also tend to 0 as $n \to \infty$, while achieving optimal prediction rates (for fixed $\alpha = 1$, $\beta = 0$), albeit assuming an infinite number of subsamples. In contrast we show optimal rates – on par with those of k-NN – for even one *denoised* subsample. This suggests, as verified experimentally, that few such denoised subsamples are required for good prediction accuracy.

The recent work of Kontorovich and Weiss (2015), of a more theoretical nature, considers a similar question as ours, and derives a *penalized* 1-NN approach shown to be statistically consistent unlike vanilla 1-NN. The approach of Kontorovich and Weiss (2015) roughly consists of finding a subsample of the data whose induced 1-NN achieves a significant margin between classes (two classes in that work). Unfortunately finding such subsample can be prohibitive (computable in time $O(n^{4.376})$) in the large data regimes of interest here. In contrast, our training phase only involves random subsamples, and cross-validation over a denoising parameter k (i.e., our training time is akin to the usual k-NN training time).

Finally, unlike in the above cited works, our rates are established for multiclass classification (for the sake of completion), and depend logarithmically on the number of classes. Furthermore, as stated earlier, our results extend beyond classification to regression, and in fact are established by first obtaining *regression* rates for estimating the so-called regression function $\mathbb{E}[Y|X]$.

Paper outline. Section 2 presents our theoretical setup and the prediction approach. Theoretical results are discussed in Section 3, and the analysis in Section 4. Experimental evaluations on real-world datasets are presented in Section 5.

2 PRELIMINARIES

2.1 Distributional Assumptions

Our main focus is classification, although our results extend to regression. Henceforth we assume we are given an i.i.d. sample $(\mathbf{X}, \mathbf{Y}) \doteq (X_i, Y_i)_1^n$ where $X \in \mathcal{X} \subset \mathbb{R}^D$, and $Y \in [L] \doteq \{1, 2, \ldots, L\}$.

The conditional distribution $P_{Y|X}$ is fully captured by the so-called *regression function*, defined as $\eta : \mathcal{X} \mapsto [0,1]^L$, where $\eta_i(x) = \mathbb{P}(Y = i|X = x)$. We assume the following on $P_{X,Y}$.

Assumption 1 (Intrinsic dimension and regularity of P_X)**.** First, for any $x \in \mathcal{X}$ and $r > 0$, define the *ball* $B(x,r) \doteq \{x' \in \mathcal{X} : \|x - x'\| \leq r\}$. We assume, there exists an integer d, and a constant C_d such that, for all $x \in \mathcal{X}, r > 0$, we have $P_X(B(x,r)) \geq C_d r^d \wedge 1$.

In this work d is unknown to the procedure. However, as is now understood from previous work (see e.g. Kpotufe (2011)), the performance of NN methods depends on such intrinsic d. We will see that the performance of the approach of interest here would also depends on such unknown d. In particular, as is argued in (Kpotufe, 2011), d is low for low-dimensional manifolds, or sparse data, so we would think of $d \ll D$ for structured data. Note that the above assumption also imposes regularity on P_X, namely by ensuring sufficient mass locally on $\mathcal{X}$ (so that NNs of a point x are not arbitrarily far from it).

Assumption 2 (Smoothness of η)**.** The function η is (λ, α)-Hölder for some $\lambda > 0, 0 < \alpha \leq 1$, i.e.,

$$\forall x, x' \in \mathcal{X}, \quad \|\eta(x) - \eta(x')\|_\infty \leq \lambda \|x - x'\|^\alpha .$$

We will use the following version of Tsybakov's noise condition (Audibert and Tsybakov, 2007), adapted to the multiclass setting.

Assumption 3 (Tsybakov noise condition)**.** For any $x \in \mathcal{X}$, let $\eta_{(l)}(x)$ denote the l'th largest element in $\{\eta_l(x)\}_{l=1}^L$. There exists $\beta > 0$, and $C_\beta > 0$ such that

$$\forall t > 0, \quad \mathbb{P}\left(\left|\eta_{(1)}(X) - \eta_{(2)}(X)\right| \leq t\right) \leq C_\beta t^\beta.$$

2.2 Classification Procedure

For any classifier $h : \mathcal{X} \mapsto [L]$, we are interested in the 0-1 classification error

$$\mathrm{err}(h) = \mathbb{P}_{X,Y}\left(h(X) \neq Y\right).$$

It is well known that the above error is mimimized by the *Bayes* classifier $h^*(x) \doteq \mathrm{argmax}_l \, \eta_l(x)$. Therefore, for any estimated classifier $\hat{h}$, we are interested in the *excess error* $\mathrm{err}(\hat{h}) - \mathrm{err}(h^*)$. We first recall the following basic nearest neighbor estimators.

Definition 1 (k-NN prediction)**.** Given $k \in \mathbb{N}$, let kNN-I(x) denote the indices of the k nearest neighbors of x in the sample $\mathbf{X}$. Assume, for simplicity, that ties are resolved so that $|k\text{NN-I}(x)| = k$. The k-NN classifier can be defined via the *regression* estimate $\hat{\eta} : \mathcal{X} \mapsto [0,1]^L$, where

$$\hat{\eta}_l(x) \doteq \frac{1}{k} \sum_{i \in k\text{NN-I}(x)} \mathbf{1}\left\{Y_i = l\right\}.$$

The k-NN classifier is then obtained as:

$$h_{\hat{\eta}}(x) = \mathrm{argmax}_l \, \hat{\eta}_l(x).$$

Finally we let $r_k(x)$ denote the distance from x to its k-th nearest neighbor.

We can now formally describe the approach considered in this work.

Definition 2 (Denoised 1-NN)**.** Consider a random subsample (**without replacement**) $\mathbf{X}'$ of $\mathbf{X}$ of size $m \leq n$. For any $x \in \mathcal{X}$, let $\mathrm{NN}(\mathbf{X}'; x)$ denote the nearest neighbor of x in $\mathbf{X}'$. The denoised 1-NN estimate at x is given as $\hat{h}(x) = h_{\hat{\eta}}(\mathrm{NN}(\mathbf{X}'; x))$, where $h_{\hat{\eta}}$ is as defined above for some fixed k.

This estimator corresponds to 1-NN over a sample $\mathbf{X}'$ where each $X'_i \in \mathbf{X}'$ is **prelabeled** as $h_{\hat{\eta}}(X'_i)$.

The resulting estimator, which we denote *subNN* for simplicity, is defined as follows.

Definition 3 (subNN)**.** Let $\{\hat{h}_i\}_{i=1}^I$, denote denoised 1-NN estimators defined over I **independent** subsamples of size m (i.e., the I sets of indices corresponding to each subsample are picked independently, although the indices in each set are picked with replacement in $[n]$). At any $x \in \mathcal{X}$, the subNN estimate $\bar{h}(x)$ is the majority label in $\{\hat{h}_i(x)\}_{i=1}^I$.

It is clear that the subNN estimate can be computed in parallel over I machines, while the final step – namely, computation of the majority vote – takes negligible time. Thus, we will view the **prediction time complexity** at any query x as the average time (over I machines) it takes to compute the 1-NN of x on each subsample. This time complexity gets better as $(m/n) \to 0$. Furthermore, we will show that, even with relatively small I (increasing variability), we can let (m/n) get small while attaining an excess error on par with that of k-NN (here $h_{\hat{\eta}}$). This is verified experimentally.

3 OVERVIEW OF RESULTS

Our main theoretical result, Theorem 1 below, concerns the statistical performance of subNN. The main technicality involves characterizing the effect of subsampling and denoising on performance. Interestingly, the rate below does not depend on the number I of subsamples: this is due to the *averaging* effect of taking majority vote accross the I submodels, and is discussed in detail in Section 4 (see proof and discussion of Lemma 2). In particular, the rate is bounded in terms of a *bad event* that is unlikely for a random submodel, and therefore unlikely to happen for a majority.

Theorem 1. *Let $0 < \delta < 1$. Let $\mathcal{V}$ denote the VC dimension of balls on $\mathcal{X}$. With probability at least $1 - L\delta$, there exists a choice of $k \in [n]$, such that the estimate $\bar{h}$ satisfies*

$$err(\bar{h}) - err(h^*) \leq C_1 \cdot \left(\frac{\mathcal{V}\ln(Ln/\delta)}{n}\right)^{\alpha(\beta+1)/(2\alpha+d)} + C_2 \cdot \left(\frac{\mathcal{V}\ln(m/\delta)}{m}\right)^{\alpha(\beta+1)/d},$$

for constants C_1, C_2 depending on $P_{X,Y}$.

The first term above is a function of the size n of the original sample, and recovers the recent optimal bounds for k-NN classification of Chaudhuri and Dasgupta (2014). We note however that the result of Chaudhuri and Dasgupta (2014) concerns binary classification, while here we consider the more general setting of multiclass. Matching lower bounds were established earlier in (Audibert and Tsybakov, 2007).

The second term, a function of the subsample size m, characterizes the additional error (over vanilla k-NN $h_{\hat{\eta}}$) due to subsampling and due to using 1-NN's at prediction time. As discussed earlier in the introduction, the first term dominates (i.e. we recover the same rates as for k-NN) whenever the subsampling ratio $(m/n) = \tilde{\Omega}(n^{-\alpha/(2\alpha+d)})$ which goes to 0 as $n \to \infty$. This is remarkable in that it suggests smaller subsample sizes are sufficient (for good accuracy) in the large sample regimes motivating the present work. We will see later that this is also supported by experiments.

As mentioned earlier, similar vanishing subsampling ratios were shown for *bagged* 1-NN in (Biau and Devroye, 2010; Biau et al., 2010; Samworth et al., 2012), but assuming an infinite number of subsamples. In contrast the above result holds for any number of subsamples, and the improvements over 1-NN are supported in experiments over varying number of subsamples, along with varying subsampling ratios.

The main technicalities and insights in establishing Theorem 1 are discussed in Section 4 below, with some proof details relegated to the appendix.

4 ANALYSIS OVERVIEW

The proof of Theorem 1 is obtained by combining the statements of Propositions 2 and 3 below. The main technicality involved is in establishing Proposition 3 which brings together the effect of noise margin β, smoothness α, and the overall error due to denoising over a subsample. We overview these supporting results in the next subsection, followed by the proof of Theorem 1.

4.1 Supporting Results

Theorem 1 relies on first establishing a rate of convergence for the k-NN regression estimate $\hat{\eta}$, used in denoising the subsamples. While such rates exist in the literature under various assumptions (see e.g. Kpotufe (2011)), we require a *high-probability* rate that holds *uniformly* over all $x \in \mathcal{X}$. This is given in Proposition 1 below, and is established for our particular setting where Y takes discrete multiclass values (i.e. $\hat{\eta}$ and η are both multivariate functions). Its proof follows standard techniques adapted to our particular aim, and is given in the appendix (supplementary material).

Proposition 1 (Uniform k-NN regression error)**.** *Let $0 < \delta < 1$. Let $\hat{\eta}$ denote the k-NN regression estimate in Definition 1. The following holds for a choice of $k = O\left(\left(\ln\frac{nL}{\delta}\right)^{\frac{d}{2\alpha+d}}(nC_d)^{\frac{2\alpha}{2\alpha+d}}\right)$. With probability at least $1 - 2\delta$ over $(\mathbf{X}, \mathbf{Y})$, we have simultaneously for all $x \in \mathcal{X}$:*

$$\|\hat{\eta}(x) - \eta(x)\|_\infty \leq C\left(\frac{\mathcal{V}\ln(nL/\delta)}{nC_d}\right)^{\frac{\alpha}{2\alpha+d}}$$

where C is a function of α, λ.

The above statement is obtained, by first remarking that, under structural assumptions on P_X, (namely that there is sizable mass everywhere locally on $\mathcal{X}$), nearest neighbor distances can be uniformly bounded with high-probability. Such nearest neighbor distances control the bias of the k-NN estimator, while its variance behaves like $O(1/k)$.

Such a uniform bound on NN distances is given in Lemma 1 below following standard insights.

Lemma 1 (Uniform bound on NN distances r_k)**.** *As in Definition 1, let $r_k(x)$ denote the distance from $x \in \mathcal{X}$ to its k'th nearest neighbor in a sample $\mathbf{X} \sim P_X^n$. Then, with probability at least $1 - \delta$ over $\mathbf{X}$, the following holds for all $k \in [n]$:*

$$\sup_{x\in\mathcal{X}} r_k(x) \leq \left(\frac{3}{C_d}\right)^{\frac{1}{d}} \cdot \max\left(\frac{k}{n}, \frac{\mathcal{V}\ln 2n + \ln\frac{8}{\delta}}{n}\right)^{\frac{1}{d}}.$$

4.2 SubNN Convergence

We are ultimately interested in the particular regression estimates induced by subsampling: the denoised 1-NN estimates $\hat{h}$ over a subsample can be viewed as $\hat{h} = \text{argmax}_{l\in[L]}\,\hat{\eta}_l^\sharp$ for a regression estimate $\hat{\eta}^\sharp(x) \doteq \hat{\eta}(\text{NN}(\mathbf{X}'; x))$, i.e., $\hat{\eta}$ evaluated at the nearest neighbor $\text{NN}(\mathbf{X}'; x)$ of x in $\mathbf{X}'$. Our first step is to relate the error of $\hat{\eta}^\sharp$ to that of $\hat{\eta}$. Here again the bound on NN distances of Lemma 1 above comes in handy since $\hat{\eta}^\sharp$ can be viewed as introducing additional bias to $\hat{\eta}$, a bias which is in turn controlled by the distance from a

query x to its NN in the subsample $\mathbf{X}'$. By the above lemma, this distance is of order $\tilde{O}(m^{-1/d})$, introducing a bias of order $\tilde{O}(m^{-\alpha/d})$ given the smoothness of η.

Thus, combining the above two results yields the following regression error on denoised estimates.

Proposition 2 (Uniform convergence of denoised 1-NN regression)**.** *Let $0 < \delta < 1$. Let $\hat{\eta}$ denote the k-NN regression estimate in Definition 1. Let $\mathbf{X}'$ denote a subsample (without replacement) of $\mathbf{X}$. Define the* denoised 1-*NN estimate $\hat{\eta}^\sharp(x) \doteq \hat{\eta}(NN(\mathbf{X}'; x))$. The following holds for a choice of $k = O\left((\mathcal{V} \ln \frac{nL}{\delta})^{\frac{d}{2\alpha+d}} (nC_d)^{\frac{2\alpha}{2\alpha+d}}\right)$. With probability at least $1 - 3\delta$ over $(\mathbf{X}, \mathbf{Y})$, and $\mathbf{X}'$, we have simultaneously for all $x \in \mathcal{X}$:*

$$\|\hat{\eta}^\sharp(x) - \eta(x)\|_\infty \leq C \left(\frac{\mathcal{V} \ln(nL/\delta)}{nC_d}\right)^{\frac{\alpha}{2\alpha+d}} + C \left(\frac{\mathcal{V} \ln(m/\delta)}{mC_d}\right)^{\frac{\alpha}{d}},$$

where C is a function of α, λ.

Proof. Define $x' = \mathrm{NN}(\mathbf{X}'; x)$ so that $\hat{\eta}^\sharp(x) = \hat{\eta}(x')$. We then have the two parts decomposition:

$$\begin{aligned}
&\|\hat{\eta}^\sharp(x) - \eta(x)\|_\infty = \|\hat{\eta}(x') - \eta(x)\|_\infty \\
&\leq \|\hat{\eta}(x') - \eta(x')\|_\infty + \|\eta(x') - \eta(x)\|_\infty \\
&\leq C \left(\frac{\mathcal{V} \ln(nL/\delta)}{nC_d}\right)^{\frac{\alpha}{2\alpha+d}} + \|\eta(x') - \eta(x)\|_\infty, \quad (1)
\end{aligned}$$

where the last inequality follows (with probability $1 - 2\delta$) from Proposition 1.

To bound the second term in inequality (1), notice that $\mathbf{X}'$ can be viewed as m i.i.d. samples from P_X. Therefore $\|x' - x\|$ can be bounded using Lemma 1. Therefore by smoothness condition on η in Assumption 2, we have with probability at least $1-\delta$, simultaneously for all $x \in \mathcal{X}$:

$$\begin{aligned}
&\|\eta(x') - \eta(x)\|_\infty \leq \lambda \|x' - x\|^\alpha \\
&\leq \lambda \left(\frac{3\mathcal{V} \ln 2m + 3 \ln \frac{8}{\delta}}{mC_d}\right)^{\frac{\alpha}{d}} \leq C \left(\frac{\mathcal{V} \ln \frac{m}{\delta}}{mC_d}\right)^{\frac{\alpha}{d}}. \quad (2)
\end{aligned}$$

Combining (1) and (2) yields the statement. □

Next we consider *aggregate regression error*, i.e., the discrepancy $(\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x))$ between the coordinates of η given by the labels $h^*(x)$ and $\bar{h}(x)$. This will be bounded in terms of the error $\phi(n, m)$ attainable by the individual denoised regression estimates (as bounded in the above proposition).

Lemma 2 (Uniform convergence of aggregate regression)**.** *Given independent subsamples $\{\mathbf{X}'_i\}_{i=1}^I$ from $(\mathbf{X}, \mathbf{Y})$, define $\hat{\eta}_i^\sharp(x) = \hat{\eta}(NN(\mathbf{X}'_i; x))$, i.e., the regression estimate $\hat{\eta}$ evaluated at the nearest neighbor $NN(\mathbf{X}'_i; x)$ of x in $\mathbf{X}'_i$. Suppose there exists $\phi = \phi(n, m)$ such that,*

$$\max_{i \in [I]} \mathbb{P}_{\mathbf{X}, \mathbf{Y}, \mathbf{X}'_i} \left(\exists x \in \mathcal{X},\ \|\hat{\eta}_i^\sharp(x) - \eta(x)\|_\infty > \phi\right) \leq \delta,$$

for some $0 < \delta < 1$. Then, let $\bar{h}$ denote the subNN estimate using subsamples $\{\mathbf{X}'_i\}_{i=1}^I$. With probability at least $1 - L\delta$ over the randomness in $(\mathbf{X}, \mathbf{Y})$ and $\{\mathbf{X}'_i\}_1^I$, the following holds simultaneously for all $x \in \mathcal{X}$:

$$\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) \leq 2\phi.$$

Remark. Notice that in the above statement, the probability of error goes from δ to $L\delta$, but does not depend on the number I of submodels. This is because of the *averaging* effect of the majority vote. For intuition, suppose B is a *bad event* and $\mathbf{1}_{B_i}$ is whether B happens for submodel i. Suppose further that $\mathbb{E}\,\mathbf{1}_{B_i} \leq \delta$ for all i. Then the likelihood of B happens for a majority of models (more than $I/2$) is

$$\mathbb{E}\,\mathbf{1}\left\{\sum_i \mathbf{1}_{B_i} \geq I/2\right\} \leq \frac{2}{I} \cdot \mathbb{E} \sum_i \mathbf{1}_{B_i} \leq 2\delta,$$

by a Markov inequality. We use this type of intuition in the proof, however over a sequence of related *bad* events, and using the fact that, the submodels' estimates are independent conditioned on $\mathbf{X}, \mathbf{Y}$.

Proof. The result is obtained by appropriately bounding the indicator $\mathbf{1}\left\{\exists x : \eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) > 2\phi\right\}$.

Let $\hat{h}_i$ denote the denoised 1-NN classifier on sample $\mathbf{X}'_i$, or for short, the i'th submodel. First notice that, if the majority vote $\bar{h}(x) \doteq l$ for some label $l \in [L]$, then at least I/L submodels $\hat{h}_i(x)$ predict l at x. In other words, we have

$$\frac{L}{I} \sum_{i=1}^I \mathbf{1}\left\{\hat{h}_i(x) = l\right\} \geq 1.$$

Therefore, fix $x \in \mathcal{X}$, and let $\bar{h}(x) \doteq l$; we then have:

$$\begin{aligned}
&\mathbf{1}\left\{\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) > 2\phi\right\} \\
&\doteq \mathbf{1}\left\{\eta_{h^*(x)}(x) - \eta_l(x) > 2\phi\right\} \\
&\leq \frac{L}{I} \sum_{i=1}^I \mathbf{1}\left\{\hat{h}_i(x) = l\right\} \mathbf{1}\left\{\eta_{h^*(x)}(x) - \eta_l(x) > 2\phi\right\} \\
&\leq \frac{L}{I} \sum_{i=1}^I \mathbf{1}\left\{\eta_{h^*(x)}(x) - \eta_{\hat{h}_i(x)}(x) > 2\phi\right\}. \quad (3)
\end{aligned}$$

We bound the above as follows. Suppose $\hat{h}_i(x) \doteq l_i$ and $h^*(x) \doteq l^*$ for some labels l_i and l^* in $[L]$. Now, if $\|\eta(x) - \hat{\eta}_i^\sharp(x)\|_\infty \leq \phi$, then $\eta_{l^*}(x) \leq \hat{\eta}_{i,l^*}^\sharp(x) + \phi$ and $\eta_{l_i}(x) \geq \hat{\eta}_{i,l_i}^\sharp(x) - \phi$. Also by definition we know that $l_i = \hat{h}_i(x)$ is the maximum entry of $\hat{\eta}_i^\sharp(x)$, so $\hat{\eta}_{i,l^*}^\sharp(x) \leq \hat{\eta}_{i,l_i}^\sharp(x)$. Therefore

$$\begin{aligned}\eta_{h^*(x)}(x) - \eta_{\hat{h}_i(x)}(x) &\leq (\hat{\eta}_{i,l^*}^\sharp(x) + \phi) - (\hat{\eta}_{i,l_i}^\sharp(x) - \phi) \\ &= (\hat{\eta}_{i,l^*}^\sharp(x) - \hat{\eta}_{i,l_i}^\sharp(x)) + 2\phi \leq 2\phi.\end{aligned}$$

In other words, $\eta_{h^*(x)}(x) - \eta_{\hat{h}_i(x)}(x) > 2\phi$ only when $\|\eta(x) - \hat{\eta}_i^\sharp(x)\|_\infty > \phi$. Thus, bound (3) to obtain:

$$\begin{aligned}&\mathbf{1}\left\{\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) > 2\phi\right\} \\ &\leq \frac{L}{I}\sum_{i=1}^{I} \mathbf{1}\left\{\|\eta(x) - \hat{\eta}_i^\sharp(x)\|_\infty > \phi\right\}.\end{aligned}$$

Finally we use the fact that, for an event $A(x)$, we have $\mathbf{1}\{\exists x \in \mathcal{X} : A(x)\} = \sup_{x\in\mathcal{X}} \mathbf{1}\{A(x)\}$. Combine this fact with the above inequality to get:

$$\begin{aligned}&\mathbf{1}\left\{\exists x \in \mathcal{X} : \eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) > 2\phi\right\} \\ &= \sup_{x\in\mathcal{X}} \mathbf{1}\left\{\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) > 2\phi\right\} \\ &\leq \sup_{x\in\mathcal{X}} \frac{L}{I}\sum_{i=1}^{I} \mathbf{1}\left\{\|\eta(x) - \hat{\eta}_i^\sharp(x)\|_\infty > \phi\right\} \\ &\leq \frac{L}{I}\sum_{i=1}^{I} \sup_{x\in\mathcal{X}} \mathbf{1}\left\{\|\eta(x) - \hat{\eta}_i^\sharp(x)\|_\infty > \phi\right\} \\ &= \frac{L}{I}\sum_{i=1}^{I} \mathbf{1}\left\{\exists x \in \mathcal{X} : \|\eta(x) - \hat{\eta}_i^\sharp(x)\|_\infty > \phi\right\}.\end{aligned}$$

The final statement is obtained by integrating both sides of the inequality over the randomness in $\mathbf{X}, \mathbf{Y}$ and the (conditionally) independent subsamples $\{\mathbf{X}_i'\}_1^I$. □

Next, Proposition 3 below states that the excess error of the subNN estimate $\bar{h}$ can be bounded in terms of the aggregate *regression* error ($\eta_{h^*} - \eta_{\bar{h}}$) considered in Lemma 2. In particular, the proposition serves to account for the effect of the *noise margin* parameter β towards obtaining faster rates than those in terms of smoothness α only.

Proposition 3. *Suppose there exists $\phi = \phi(n, m)$ such that, with probability at least $1 - \delta$ over the randomness in $(\mathbf{X}, \mathbf{Y})$ and the subsamples $\{\mathbf{X}_i'\}_{i=1}^I$, we have simultaneously for all $x \in \mathcal{X}$, $\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) < 2\phi$.*

Then with probability at least $1 - \delta$, the excess classification error of the estimate $\bar{h}$ satisfies:

$$err(\bar{h}) - err(h^*) \leq C_\beta(2\phi)^{\beta+1}.$$

Proof. Since, for any classifier h, $\eta_{h(x)}(x) \doteq \mathbb{P}(Y = h(x))$, the excess error of the sub-NN classifier $\bar{h}$ can be written as $\mathbb{E}_X\left[\eta_{h^*(X)}(X) - \eta_{\bar{h}(X)}(X)\right]$.

Thus, the assumption in the proposition statement – that for all $x \in X$, $\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) < 2\phi$ (with probability at least $1 - \delta$) – yields a trivial bound of 2ϕ on the excess error. We want to refine this bound.

Let $\eta_{(l)}(x)$ be the l-th largest entry in the vector $\eta(x) = \{\eta_l\}_{l=1}^L$, and define $\Delta(x) \doteq \eta_{(1)}(x) - \eta_{(2)}(x)$.

Then at any fixed point x, we can refine the bound on excess error at x (namely $\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x)$), by separately considering the following exhaustive conditions $A(x)$ and $B(x)$ on x:

- A: $\Delta(x) \geq 2\phi$, in which case the excess error is 0. This follows from $\eta_{(1)}(x) = \eta_{h^*(x)}(x)$ and that:

  $$\eta_{h^*(x)}(x) - \eta_{\bar{h}(x)}(x) < 2\phi \leq \eta_{h^*(x)}(x) - \eta_{(2)}(x)$$

  In other words $\eta_{\bar{h}(x)}(x)$ is larger than $\eta_{(2)}$, so equals $\eta_{(1)}(x) \doteq \eta_{h*(x)}(x)$.
- B: $\Delta(x) < 2\phi$, in which case the excess error cannot be refined at x. However, the total mass of such x's is at most $C_\beta(2\phi)^\beta$ by Tsybakov's noise condition (Assumption 3).

Combining these conditions, we have with probability at least $1 - \delta$ that the excess error satisfies:

$$\begin{aligned}\mathrm{err}(\bar{h}) - \mathrm{err}(h^*) &= \mathbb{E}_X\left[\eta_{h^*(X)}(X) - \eta_{\bar{h}(X)}(X)\right] \\ &\leq \mathbb{E}_X\left[0 \cdot \mathbf{1}\{A(X)\} + 2\phi \cdot \mathbf{1}\{B(X)\}\right] \\ &\leq 2\phi \cdot \mathbb{E}\left[\mathbf{1}\{B(X)\}\right] \leq 2\phi \cdot \mathbb{P}(\Delta(x) \leq 2\phi) \\ &\leq C_\beta(2\phi)^{\beta+1}.\end{aligned}$$

□

Combining the results of this section yield the main theorem whose proof is given next.

4.3 Proof Of Theorem 1

Our main result follows easily from Propositions 2 and 3. This is given below.

Proof. Fix any $0 < \delta < 1$. Note that the conditions of Lemma 2 are verified in Proposition 2, namely that, with probability $1 - 3\delta$, all regression errors (of submodels) are bounded by

$$\phi \doteq C\left(\frac{\mathcal{V}\ln(Ln/\delta)}{n}\right)^{\alpha/(2\alpha+d)} + C\left(\frac{\mathcal{V}\ln(m/\delta)}{m}\right)^{\alpha/d}, \tag{4}$$

Table 1: Datasets Used In Evaluating subNN

Name	#train	#test	#dimension	#classes	Description
MiniBooNE	120k	10k	50	2	Particle identification (Roe, 2010)
TwitterBuzz	130k	10k	50	2	Buzz in social media (François Kawala, 2013)
LetterBNG	34k	10k	16	26	English alphabet (ML)
NewsGroups20	11k	7.5k	130k	20	Document classification (Mitchell, 1999)
YearPredMSD	34k	10k	90	regression	Release year of songs (Bertin-Mahieux, 2011)
WineQuality	5.5k	1.0k	12	regression	Quality of wine (Cortez, 2009)

where C is a constant depending on α and λ.

Next, the conditions of Proposition 3 are obtained in Lemma 2 with the same setting of ϕ. It follows that with probability at least $1 - 3L\delta$, we have:

$$\text{err}(\bar{h}) - \text{err}(h^*) \leq C_\beta (2\phi)^{\beta+1},$$

with ϕ as in (4). Given that $g(x) = x^{\beta+1}$ is a convex function, we conclude by applying Jensen's inequality (viewing $(1/2C) \cdot \phi$ as an average of two terms). □

5 EXPERIMENTS

Experimental Setup. Data is standardized along each coordinate of X.

- *Fitting subNN.* We view the subsample size m and the number I of subsamples as exogenous parameters determined by the practical constraints of a given application domain. Namely, smaller m yields faster prediction and is driven by prediction time requirement, while larger I improves prediction error but is constrained by available computing units. However, much of our experiments concern the sensitivity of subNN to m and I, and yield clear insights into tradeoffs on these choices. Thus, for any fixed choice of m and I, we choose k by 2-fold cross-validation. The search for k is done in two stages: first, the best value k' (minimizing validation error) is picked on a log-scale $\{2^i\}_{i=1}^{\lceil \log n \rceil}$, then a final choice for k is made on a refined linear range $[\lceil k'/2 \rceil - 10 : 2k' + 10]$.
- *Fitting k-NN.* The choice of k for vanilla k-NN is also made in two stages as above.

Table 1 describes the datasets used in the experiments. We use k-d-tree for fast NN search from Python `scikit-learn` for all procedures and all datasets except NewsGroups20, for which we perform a direct search (due to high-dimensionality and sparsity). As explained earlier, our main focus is on classification, however theoretical insights from previous sections extend to regression, as substantiated in this section.

Results. Our main experimental results are described in Table 2, showing the relative errors (error of the method divided by that of vanilla k-NN) and relative prediction time (prediction time divided by that of k-NN) for versions of subNN((m/n), I), where m/n is the subsampling ratio used, and I is the number of subsamples. For regression datasets, the error is the MSE, while for classification we use 0-1 error. The prediction time reported for the subNN methods is the maximum time over the I subsamples plus aggregation time, reflecting the effective prediction time for the distributed-computing settings motivating this work. The results support our theoretical insights, namely that subNN can achieve accuracy close or matching that of k-NN, while at the same time achieving fast prediction time, on par or better than those of 1-NN.

- *Sensitivity to m and I.* As expected, better times are achievable with smaller subsample sizes, while better prediction accuracy is achievable as more subsamples tend to reduce variability. This is further illustrated for instance in Figure 1, where we vary the number of subsamples. Interestingly in this figure, for the MiniBoone dataset, the larger subsampling ratio 0.75 yields the best accuracy over any number of subsamples, but the gap essentially disappears when enough subsamples are used. We thus have the following prescription in choosing m and I: while small values of I work generally well, large values of I can only improve accuracy; on the other hand, subsampling ratios of 0.1 yield good time-accuracy tradeoffs accross datasets.

- *Benefits of denoising.* In Figure 2, we compare subNN with pure bagging of 1-NN models. As suggested by theory, we see that the bagging approach does indeed require considerably more subsamples to significantly improve over the error of vanilla 1-NN. In contrast, the accuracy of subNN quickly tends to that of k-NN, in particular for TwitterBuzz where 1 or 3 subsamples are sufficient to statistically close the gap, even for small subsampling ratio. This could be due to hidden but beneficial structural aspects of this data. In all cases, the experiments further highlights the benefits of our simple denoising step, as a variance reduction technique. This is further supported by the error-bars (std) over 5 repetitions as shown in Figure 2.

Conclusion. We propose a procedure with theoretical guarantees, which is easy to implement over distributed computing resources, and which achieves good time-accuracy tradeoffs for nearest neighbor methods.

Table 2: Ratios Of *Error Rates* and *Prediction Times* Over Corresponding Errors And Times Of k-NN.

	Relative Error			Relative Time		
Data	1NN	subNN(0.1,10)	subNN(0.75,10)	1NN	subNN(0.1,10)	subNN(0.75,10)
MiniBooNE	1.280	1.039	**1.027**	0.609	**0.247**	0.547
TwitterBuzz	1.405	**1.000**	1.005	0.550	**0.185**	0.522
LetterBNG	1.127	**1.086**	1.144	0.459	**0.219**	0.396
NewsGroups20	1.122	1.206	**1.002**	0.610	**0.081**	0.668
YearPredMSD	1.859	**1.082**	1.110	0.847	**0.025**	0.249
WineQuality	1.276	**1.011**	1.018	0.989	**0.885**	0.906

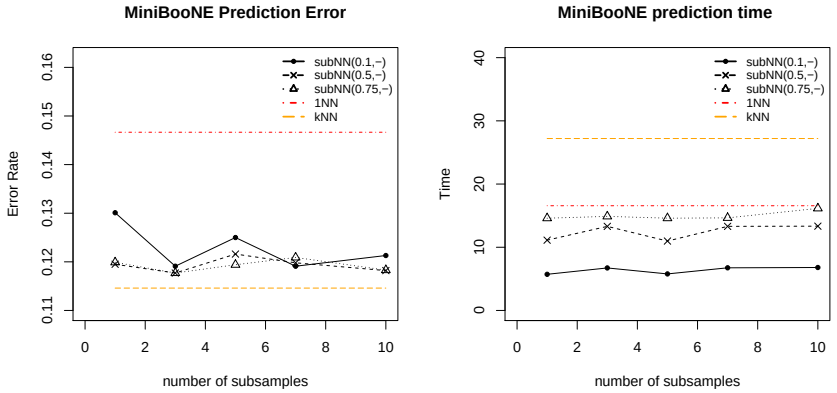

Figure 1: Comparing the Effect Of Subsampling Ratios on Prediction And Time Performance Of SubNN. Shown are SubNN estimates using subsampling ratios 0.1, 0.5, and 0.75.

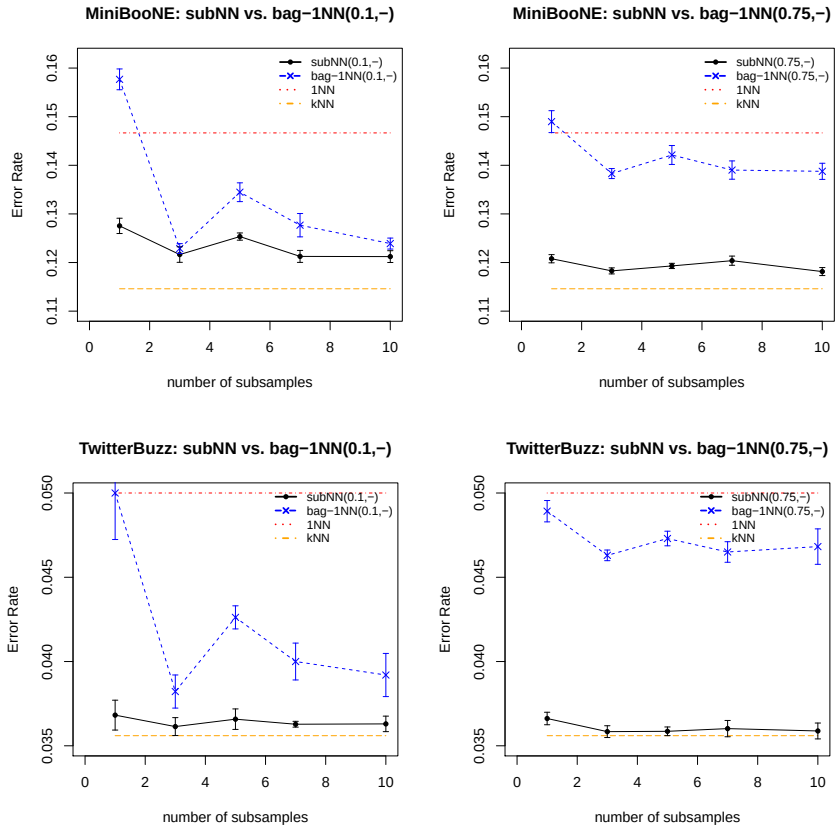

Figure 2: Bagged 1NN Compared With SubNN Using Subsampling Ratios 0.1 (Left) And 0.75 (Right).

References

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

T. Bertin-Mahieux. Yearpredictionmsd data set. `https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD`, 2011.

A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbors. *International Conference on Machine Learning (ICML)*, 2006.

Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10): 2499–2518, 2010.

Gérard Biau, Frédéric Cérou, and Arnaud Guyader. On the rate of convergence of the bagged nearest neighbor estimate. *Journal of Machine Learning Research*, 11 (Feb):687–712, 2010.

Timothy I Cannings, Thomas B Berrett, and Richard J Samworth. Local nearest neighbour classification with applications to semi-supervised learning. *arXiv preprint arXiv:1704.00642*, 2017.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

Kenneth L. Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, 2005.

Paulo Cortez. Wine quality data set. `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`, 2009.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, pages 1371–1385, 1994.

Evelyn Fix and Joseph L Hodges Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, DTIC Document, 1951.

et. al. François Kawala. Buzz in social media data set. `https://archive.ics.uci.edu/ml/datasets/Buzz+in+social+media+`, 2013.

Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *VLDB*, volume 99, pages 518–529, 1999.

Aryeh Kontorovich and Roi Weiss. A bayes consistent 1-nn classifier. In *AISTATS*, 2015.

S. Kpotufe. k-NN Regression Adapts to Local Intrinsic Dimension. *Advances in Neural Information Processing Systems (NIPS)*, 2011.

R. Krauthgamer and J. Lee. Navigating nets: Simple algorithms for proximity search. *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2004.

Sanjeev R Kulkarni and Steven E Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41 (4):1028–1039, 1995.

Tom Mitchell. Twenty newsgroups data set. `https://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups`, 1999.

Open ML. Open ml machine learning platform. `https://www.openml.org/d/1378`.

Amit Moscovich, Ariel Jaffe, and Boaz Nadler. Minimax-optimal semi-supervised regression on unknown manifolds. *arXiv preprint arXiv:1611.02221*, 2016.

Byron Roe. Miniboone particle identification data set. `https://archive.ics.uci.edu/ml/datasets/MiniBooNE+particle+identification`, 2010.

Richard J Samworth et al. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.