Efficient Classification with Adaptive KNN

Puning Zhao and Lifeng Lai
University of California Davis
{pnzhao, lflai}@ucdavis.edu

Abstract

In this paper, we propose an adaptive kNN method for classification, in which different k are selected for different test samples. Our selection rule is easy to implement since it is completely adaptive and does not require any knowledge of the underlying distribution. The convergence rate of the risk of this classifier to the Bayes risk is shown to be minimax optimal for various settings. Moreover, under some special assumptions, the convergence rate is especially fast and does not decay with the increase of dimensionality.

1 Introduction

k Nearest Neighbor (kNN) method is a simple and popular approach to nonparametric classification. In this setup, we have N identical and independently distributed (i.i.d) training samples $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, N$, which are all drawn from a pair of random variables $(\mathbf{X}, Y)$ with an unknown distribution. Given any new test sample $\mathbf{x}$, the kNN classifier assigns label Y that is determined by the majority vote of the k nearest neighbors of this new sample among the training dataset (Fix 1951). In the standard kNN method, k is fixed for all test samples. It has been proved that if k grows with N and k/N goes to zero, then as the sample size N increases, the risk of kNN classifier converges to the Bayes risk, which is defined as the minimum possible error probability among all classifiers (Cover and Hart 1967; Biau and Devroye 2015; Stone 1977; Devroye et al. 1994; Devroye, Györfi, and Lugosi 2013; Cérou and Guyader 2006). To maximize the convergence rate, the growth rate of k needs to be properly selected to achieve a desirable bias and variance tradeoff. It has been shown that, if the support set of the distribution of the feature vector $\mathbf{X}$ is bounded and the probability density function (pdf) of $\mathbf{X}$ is bounded away from zero, then the risk of the standard kNN classifier converges to the Bayes risk with the best rate among all classifiers in the minimax sense, if the growth rate of k is properly selected (Audibert 2004; Kohler and Krzyzak 2007; Györfi 1981; Audibert and Tsybakov 2007; Chaudhuri and Dasgupta 2014; Döring, Györfi, and Walk 2017; Mammen and Tsybakov 1999). On the contrary, if the distribution of $\mathbf{X}$ has tails, then the convergence rate of the standard kNN classifier is no longer minimax optimal (Gadat, Klein, and Marteau 2016; Zhao and Lai 2019). This can be explained by the fact that the bias and the variance of the prediction vary among different locations, and hence the optimal k that achieves the best bias and variance tradeoff also depends on the feature vector. If we use the same k for all test samples, then the selected k may not be universally optimal. Therefore, to improve the performance of kNN classifier, it is necessary to design a rule such that k is selected adaptively based on the locations of test samples (Ougiaroglou et al. 2007; Sun and Huang 2010; Balsubramani et al. 2019; Kpotufe 2011).

Several adaptive kNN classifiers have been designed and analyzed in (Gadat, Klein, and Marteau 2016; Cannings, Berrett, and Samworth 2017; Zhao and Lai 2019). In (Gadat, Klein, and Marteau 2016), a 'sliced nearest neighbor' method was proposed. This method divides the whole support into several regions depending on the pdf of $\mathbf{X}$, and uses different k in different regions. Moreover, it is shown in (Gadat, Klein, and Marteau 2016) that this adaptive kNN method is minimax rate optimal for distributions with tails. However, this classifier requires the precise knowledge of the pdf $f(\mathbf{x})$, which is usually unavailable. If we use an estimate of pdf $\hat{f}$, then the theoretical guarantee of this classifier has not been established after we take the estimation error into consideration. (Cannings, Berrett, and Samworth 2017) proposed a method for semi-supervised learning, which means that, apart from the labeled training samples, we also have a much larger set of unlabeled samples. The pdf $f(\mathbf{x})$ can be estimated using these unlabeled samples, and then the optimal k can be selected based on the estimated pdf. A new adaptive kNN method was proposed in (Zhao and Lai 2019), which relies entirely on the training dataset, without requiring the precise knowledge of the pdf $f(\mathbf{x})$, or a large number of unlabeled samples. In particular, for each test sample, the adaptive kNN method in (Zhao and Lai 2019) selects k according to the number of training samples that fall in a ball centered at the test sample with a fixed radius, so that one can achieve a desirable bias and variance tradeoff. Moreover, theoretical analysis shows that this method is minimax rate optimal for a broad range of distributions, under some tail, smoothness and margin assumptions. However, there are some design parameters of this method that need to be carefully tuned, and the optimal values of these parameters depend on the properties of the underlying joint distributions of the feature $\mathbf{X}$ and the label Y. Therefore, the implementation of this method still

relies partly on the underlying distribution, which is unknown in practice.

In this paper, we propose a new adaptive kNN classifier that does not need any prior knowledge of the underlying distribution. We consider binary classification with Y taking values in $\{-1, 1\}$, and define $\eta(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ as the conditional expectation of the label. The basic idea of our new method is to let k grow step by step. In each step, we estimate $\eta(\mathbf{x})$ with the average of labels of k nearest neighbors of $\mathbf{x}$. If the estimated $\eta(\mathbf{x})$ is sufficiently far away from zero (will be made precise in the sequel), then we can believe with high confidence that the sign of the estimated $\eta(\mathbf{x})$ is correct, and hence the prediction will be the same as that of the Bayes classifier. In this case, our algorithm returns a prediction using the sign of the current estimate of $\eta(\mathbf{x})$. On the contrary, if $\eta(\mathbf{x})$ is not large enough, then the algorithm lets k increase, until the random error bound is lower than the estimated $|\eta(\mathbf{x})|$, or k reaches an upper bound k_{max}. The key difference between our method and previous adaptive kNN methods in (Gadat, Klein, and Marteau 2016; Cannings, Berrett, and Samworth 2017; Zhao and Lai 2019) is that previous methods select k based on the real or estimated pdf $f(\mathbf{x})$, while our method selects k based on the estimated $\eta(\mathbf{x})$. We then analyze the convergence property of this new method. To begin with, we establish a general convergence bound, which depends on the joint distribution of $\mathbf{X}$ and Y, and holds universally without any assumptions. Based on such a general bound, we then obtain convergence rates for some common classes of distributions, as were discussed in (Chaudhuri and Dasgupta 2014) and (Zhao and Lai 2019). Our results show that, for both distributions with or without tails, the proposed method is minimax rate optimal. Furthermore, we show that under a special case where the Bayes boundary is linear and some other assumptions hold, the convergence rate of the new adaptive kNN classifier is fast and does not become worse as the dimensionality increases.

Compared with the existing adpative kNN methods (Gadat, Klein, and Marteau 2016; Cannings, Berrett, and Samworth 2017; Zhao and Lai 2019), our new method has the following advantages. Firstly, the new method does not require any prior knowledge of the underlying distribution, and the parameter tuning is convenient. The only parameter of this new method is k_{max}, the largest value of k that the algorithm will use. Unlike previous methods, in which the parameters need to be carefully tuned to achieve the optimal bias and variance trade-off, in this new method, it is always safe to use a larger k_{max}, although sometimes it is sufficient to use a smaller k_{max} to reduce the computational complexity without significantly deteriorating the performance. Moreover, the performance of our new method is also competitive. Despite that previous methods in (Gadat, Klein, and Marteau 2016; Zhao and Lai 2019) are already minimax rate optimal under some tail, margin and smoothness assumptions, the convergence rate can still be further improved, since the minimax lower bound is only tight for the worst case among all distributions that satisfy those assumptions. For many common distributions, it is possible to achieve a faster convergence rate. We show that the convergence rate of our new classifier is usually better than the minimax bound for many common distributions.

Our method is a simplified form of the method proposed in (Balsubramani et al. 2019). In (Balsubramani et al. 2019), there are several parameters to be tuned. On the contrary, our method has only one design parameter k_{max}, and the value of this parameter is not crucial, hence our method is easier to use. Moreover, we analyze the convergence rate of the overall excess risk of the adaptive kNN method for a broad class of different cases, while (Balsubramani et al. 2019) only analyze its local convergence.

The remainder of this paper is organized as follows. In Section 2, we describe the detailed design of our new proposed method. We then derive a general convergence bound of the proposed method without any assumptions in Section 3. In Section 4, we analyze the convergence rate under certain common assumptions. In Section 5, we show that for a special case, the new method has a fast convergence rate, and this rate does not become worse as the dimensionality increases. Finally, numerical examples and the concluding remarks are provided in Section 6 and 7, respectively.

2 The Proposed Method

In this section, we describe the proposed adaptive kNN classifer. Let the feature vector $\mathbf{X}$ and target Y take values in $\mathbb{R}^d$ and $\{-1, 1\}$, respectively. $(\mathbf{X}, Y)$ is a pair of random variables that follows an unknown distribution. The training dataset contains N samples $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, N$, which are i.i.d drawn from this joint distribution. Based on these training samples, our goal is to learn a function $g : \mathbb{R}^d \to \{-1, 1\}$, which can be used to make a prediction $\hat{Y}$. In this paper, we use $0 - 1$ loss function to evaluate the quality of classification, i.e.

$$L(\hat{Y}, Y) = \begin{cases} 0 & \text{if} \quad \hat{Y} = Y \\ 1 & \text{if} \quad \hat{Y} \neq Y. \end{cases} \tag{1}$$

With this loss function, the risk of a classifier is defined as

$$R(g) = \mathbb{E}[L(\hat{Y}, Y)] = \mathrm{P}(g(\mathbf{X}) \neq Y). \tag{2}$$

Moreover, define the regression function as

$$\eta(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]. \tag{3}$$

From (1), it can be shown that

$$\eta(\mathbf{x}) = \mathrm{P}(Y = 1|\mathbf{X} = \mathbf{x}) - \mathrm{P}(Y = -1|\mathbf{X} = \mathbf{x}). \tag{4}$$

It can be shown (Chaudhuri and Dasgupta 2014; Döring, Györfi, and Walk 2017) that the optimal classification rule is $g^*(\mathbf{x}) = \mathrm{sign}(\eta(\mathbf{x}))$, since if $\eta(\mathbf{x}) > 0$, then Y is more likely to be 1, thus predicting $Y = 1$ minimizes the error probability, and vice versa. The corresponding risk, called Bayes risk, is

$$R^* = \mathrm{P}(g^*(\mathbf{X}) \neq Y)) = \mathbb{E}\left[\frac{1 - |\eta(\mathbf{X})|}{2}\right]. \tag{5}$$

In practice, η is unknown, and hence $g^*(\mathbf{x})$ is also unknown. The kNN classification rule returns $g(\mathbf{x}) = \mathrm{sign}(\hat{\eta}_k(\mathbf{x}))$ instead, in which $\hat{\eta}_k(\mathbf{x})$ is the estimated regression function, based on the average of the labels of k nearest neighbors of $\mathbf{x}$, i.e.

$$\hat{\eta}_k(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} Y^{(i)}, \tag{6}$$

with $Y^{(i)}$ being the label of the i-th nearest neighbor of $\mathbf{x}$.

In the standard kNN classifer, the value of k remains the same regardless of the value of $\mathbf{x}$. In this paper, we design a new adaptive kNN classifier in which the value of k is different for different $\mathbf{x}$. Our design is motivated by the following intuitions. The prediction of a kNN classifier is the same as the Bayes classifier if $g(\mathbf{x}) = g^*(\mathbf{x})$, thus the best value of k maximizes the probability that $\hat{\eta}_k(\mathbf{x})$ and $\eta(\mathbf{x})$ have the same signs. Such an optimal k changes with $\mathbf{x}$. For example, consider two test samples located at $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively, with $f(\mathbf{x}_1) > f(\mathbf{x}_2)$ and $|\eta(\mathbf{x}_1)| < |\eta(\mathbf{x}_2)|$. In this case, the optimal k of the first sample is larger than that of the second one. For the first sample, since $f(\mathbf{x}_1)$ is larger, the kNN distances are smaller, thus we can use a larger k without worrying too much about the bias. On the other hand, since $|\eta(\mathbf{x}_1)|$ is relatively small, it is necessary to use a large k to reduce the estimation variance, so that the label of $\eta(\mathbf{x})$ can be more likely to be inferred correctly. On the contrary, for the second sample, a smaller k is better. In the standard kNN method, k is fixed for all samples, therefore it is inevitable that k is suboptimal for some test points. As a result, in order to improve the classification accuracy, we need to estimate $f(\mathbf{x})$ or $\eta(\mathbf{x})$ to help us decide the best k to use.

Previous adaptive kNN methods in (Gadat, Klein, and Marteau 2016; Cannings, Berrett, and Samworth 2017; Zhao and Lai 2019) solve this problem by selecting k based on the real or estimated pdf $f(\mathbf{x})$, so that larger k is used where $f(\mathbf{x})$ is high, and vice versa. Our new method takes a different approach. We select k based on estimated $|\eta(\mathbf{x})|$ instead of $f(\mathbf{x})$. The detailed algorithm is shown as Algorithm 1.

Algorithm 1 Adaptive kNN classification algorithm

input Training samples $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_N, Y_N)$; test point $\mathbf{x}$; k_{max}
output Prediction $\hat{Y}$
  Find the labels $Y^{(i)}$, $i = 1, \ldots, k_{max}$ of the nearest neighbors of $\mathbf{x}$
  $k \leftarrow \lceil \ln^2 N \rceil$
  $\hat{\eta} \leftarrow (1/k) \sum_{i=1}^{k} Y^{(k)}$
  while $k \leq k_{max}$ **do**
    if $|\hat{\eta}| > k^{-1/2} \ln N$ **then**
      return $\text{sign}(\hat{\eta})$
    else
      $\hat{\eta} \leftarrow \eta + \frac{1}{k+1}(Y^{(k+1)} - \hat{\eta})$
      $k \leftarrow k + 1$
    end if
  end while
  return a random value from $\{-1, 1\}$

The main idea of this adaptive kNN algorithm is to let k grow step by step. In each step, we calculate the estimated regression function at the test point, i.e. $\hat{\eta}_k(\mathbf{x})$, based on the current k. Considering that the variance of $\hat{\eta}_k(\mathbf{x})$ scales with $1/k$, if $|\hat{\eta}_k(\mathbf{x})|$ is larger than $k^{-1/2} \ln N$, then with high probability the random error will not change the sign of $\hat{\eta}_k(\mathbf{x})$. However, it is still possible that the sign of $\hat{\eta}_k(\mathbf{x})$ is not correct, due to the estimation bias, which increases with the kNN distances. Therefore, in our algorithm, we use the smallest k such that

$$|\hat{\eta}_k(\mathbf{x})| > k^{-1/2} \ln N,$$

in order to control the bias. On the other hand, if $|\hat{\eta}_k(\mathbf{x})|$ is not large enough, then it is likely that the sign is not correct due to the random error, therefore we continue to increase k, until k reaches an upper bound k_{max}, which is the only design parameter of our method. If $|\hat{\eta}_k(\mathbf{x})| \leq k^{-1/2} \ln N$ for all $k \leq k_{max}$, then the determination of the sign of $\eta(\mathbf{x})$ is hard, since $|\eta(\mathbf{x})|$ is too low. In this case, we give up the prediction and return a random result. In Algorithm 1, k starts from $\lceil \ln^2 N \rceil$, since if $k < \ln^2 N$, then $k^{-1/2} \ln N > 1$, which will always lead to the increase of k, thus it is not necessary to try with these small k values. Similar idea has been proposed in (Balsubramani et al. 2019). Compared with (Balsubramani et al. 2019), our proposed method is simpler, since we are using less parameters. Moreover, we derive bounds of the overall excess risk for a broad class of different f and η functions, while (Balsubramani et al. 2019) only shows the local convergence rate.

In previous adaptive kNN methods (Zhao and Lai 2019; Gadat, Klein, and Marteau 2016), there are parameters that need to be tuned carefully to achieve the minimax optimal rate, and the optimal parameter depends on the property of the distribution, such as tail, margin and smoothness parameters. However, these properties are usually unknown, therefore it is hard to find the optimal parameters. Our new method can solve this problem, since this method has only one design parameter k_{max}, and we do not need to tune it very carefully. In fact, in Sections 3 and 4, we show that using a large k_{max} can always achieve good accuracy, although sometimes we can use a smaller k_{max} to accelerate the computation.

Note that compared with the standard kNN method, our new method does not increase the time complexity of computation. In Algorithm 1, the computation cost includes nearest neighbor search and the update of η. The updating requires $\mathcal{O}(k_{max})$ time for each test sample, while the nearest neighbor search requires more time (Bentley, Stanat, and Williams Jr 1977; Leibe, Mikolajczyk, and Schiele 2006). As a result, the overall time complexity of our new method with parameter k_{max} is the same as that of standard kNN method with $k = k_{max}$.

3 General Bound

In this section, we show a general convergence bound of this new adaptive kNN classifier without any restricting assumptions. For any set $S \in \mathbb{R}^d$, define

$$\eta(S) := \mathbb{E}[Y|\mathbf{X} \in S], \tag{7}$$

which is the weighted average of the regression function within S. Based on the regression function, the excess risk of a classifier $g(\mathbf{x})$ is

$$R - R^* = \mathbb{E}[|\eta(\mathbf{X})|\mathbf{1}(g(\mathbf{X}) \neq \text{sign}(\eta(\mathbf{X})))]. \tag{8}$$

Denote $B(\mathbf{x}, r) = \{\mathbf{x}' \in \mathbb{R}^d | \, \|\mathbf{x}' - \mathbf{x}\| < r\}$ as the ball

centering at $\mathbf{x}$ with radius r. Define

$$\Delta(\mathbf{x}, r) := \begin{cases} \sup_{r'<r} (\eta(\mathbf{x}) - \eta(B(\mathbf{x}, r'))) & \text{if} \quad \eta(\mathbf{x}) \geq 0 \\ \sup_{r'<r} (\eta(B(\mathbf{x}, r')) - \eta(\mathbf{x})) & \text{if} \quad \eta(\mathbf{x}) < 0, \end{cases} \tag{9}$$

$$\rho(\mathbf{x}) := \sup \left\{ r | \Delta(\mathbf{x}, r) \leq \frac{1}{2} |\eta(\mathbf{x})| \right\}, \tag{10}$$

$$p(\mathbf{x}) := \mathrm{P}(B(\mathbf{x}, \rho(\mathbf{x}))). \tag{11}$$

Then the convergence rate of the adaptive kNN classification algorithm is shown in the following theorem.

Theorem 1 *Define*

$$\begin{aligned} S_N &= \{\mathbf{x} \,||\eta(\mathbf{x})| \\ &\leq \max \left\{ \sqrt{\frac{32}{Np(\mathbf{x})}} \ln N, 4 \ln N k_{max}^{-\frac{1}{2}} \right\} \Bigg\}, \end{aligned} \tag{12}$$

then the excess risk is bounded by

$$R - R^* \leq \int_{S_N} f(\mathbf{x}) |\eta(\mathbf{x})| d\mathbf{x} + 2(N + 20) e^{-2 \ln^2 N}. \tag{13}$$

Intuitively, S_N in (12) is the region in which the classification is hard, which means that for any k, we can not ensure that the inference of the sign of $\eta(\mathbf{x})$ is correct with high probability, since $|\eta(\mathbf{x})|$ is too small, and thus the estimation error can make $\hat{\eta}(\mathbf{x})$ has different sign with $\eta(\mathbf{x})$. On the contrary, for all $\mathbf{x} \notin S_N$, there exists a k such that the prediction is the same as that of the Bayes classifier with a high probability. Such a k value can be found using Algorithm 1. From (12), we observe that S_N shrinks with N if $k_{max} / \ln^2 N \to \infty$, and its limit is $\{\mathbf{x} | \eta(\mathbf{x}) = 0\}$. Therefore, the first term in the right hand side of (13) converges to zero. The second term in (13) converges faster than any polynomial of N. If $\eta(\mathbf{x})$ crosses zero in its support, then the first term usually dominates, since the size of S_N shrinks polynomially with N.

4 Convergence Rate with Assumptions

In this section, we derive the convergence rate of the new kNN classifier under some common assumptions that have been analyzed in (Chaudhuri and Dasgupta 2014; Zhao and Lai 2019). To begin with, we analyze the convergence rate under certain margin and probabilistic continuous assumptions. These assumptions are proposed in (Chaudhuri and Dasgupta 2014), and are usually satisfied if the feature distribution has bounded support and the pdf is bounded away from zero. We then analyze the convergence rate of the proposed method for distributions with tails under some margin, smoothness and tail assumptions. These assumptions were used in (Zhao and Lai 2019) to analyze distributions with tails. For both cases, we show that the new kNN classifier is minimax rate optimal up to a log polynomial factor, as long as k_{max} grows sufficiently fast with N.

Convergence Rate under Margin and Probabilistic Continuous Assumptions

We first analyze the convergence rate of the new adaptive kNN classifier under certain margin and probabilistic continuous assumptions. These assumptions have been used in (Chaudhuri and Dasgupta 2014) for deriving the bound of the standard kNN classification.

Theorem 2 *Assume $f(\mathbf{x})$ and $\eta(\mathbf{x})$ satisfy the following assumptions:*

(a) There exist two constants C_a, α such that

$$P(0 < |\eta(\mathbf{X})| < t) \leq C_a t^\alpha \tag{14}$$

for all $t > 0$;

(b) There exist two constants L, γ such that

$$|\eta(B(\mathbf{x}, r)) - \eta(\mathbf{x})| \leq L P^\gamma(B(\mathbf{x}, r)) \tag{15}$$

for all $r > 0$, in which $P(B(\mathbf{x}, r))$ is the probability mass of $B(\mathbf{x}, r)$.

The convergence rate of the excess risk of the adaptive kNN classifier is bounded by:

$$\begin{aligned} R - R^* &= \mathcal{O}\left(N^{-\frac{\gamma(1+\alpha)}{2\gamma+1}} (\ln N)^{\frac{2\gamma(1+\alpha)}{2\gamma+1}} \right. \\ &\quad \left. + k_{max}^{-\frac{1}{2}(1+\alpha)} (\ln N)^{1+\alpha} \right). \end{aligned} \tag{16}$$

If $k_{max} \gtrsim N^{\frac{2\gamma}{2\gamma+1}}$, then from (16),

$$R - R^* = \mathcal{O}\left(N^{-\frac{\gamma(1+\alpha)}{2\gamma+1}} (\ln N)^{\frac{2\gamma(1+\alpha)}{2\gamma+1}} \right). \tag{17}$$

The convergence rate in (16) matches the previous results in (Chaudhuri and Dasgupta 2014; Kohler and Krzyzak 2007), as well as the minimax lower bound in (Audibert and Tsybakov 2007), up to a log-polynomial factor. Assumption (a) is common in many previous works (Kohler and Krzyzak 2007; Döring, Györfi, and Walk 2017; Chaudhuri and Dasgupta 2014; Gadat, Klein, and Marteau 2016), which restricts the noise level. The convergence rate is faster with a larger α, since misclassification is easier to occur where $\eta(\mathbf{x})$ is close to zero. Assumption (b) requires η to be continuous. Although assumption (b) does not strictly require that the pdf is bounded away from zero, it can be observed that if the distribution has tails, then at the region with low pdf, $\mathrm{P}(B(\mathbf{x}, r))$ is small, and thus $|\eta(B(\mathbf{x}, r)) - \eta(\mathbf{x})|$ need to be very close to zero, which can be too restrictive. Therefore, this assumption usually holds for the cases such that the feature distribution has bounded support and the pdf $f(\mathbf{x})$ is bounded away from zero.

Convergence Rate for Distributions with Tails

We now analyze the proposed adaptive kNN classifier for pdfs that can be arbitrarily close to zero. In particular, we analyze the convergence rate under the same assumptions as in (Zhao and Lai 2019).

Theorem 3 *Assume that there exist constants C_a, C_b, C_c, C_d, α, β and D, such that*

(a) For any $t > 0$,

$$P(0 < |\eta(\mathbf{X})| < t) \leq C_a t^\alpha;$$

(b) For any $t > 0$,

$$P(f(\mathbf{X}) < t) \leq C_b t^\beta;$$

(c) For any $r > 0$,

$$|\eta(B(\mathbf{x}, r)) - \eta(\mathbf{x})| \leq C_c r^2;$$

(d) For any $r \in (0, D]$,

$$P(B(\mathbf{x}, r)) \geq C_d f(\mathbf{x}) r^d,$$

then

$$R - R^* = \mathcal{O}\left(N^{-\min\{\beta, \frac{2\beta(1+\alpha)}{\beta d+2(\alpha+2\beta)}\}} \ln^c N + k_{max}^{-\frac{1}{2}(1+\alpha)} (\ln N)^{1+\alpha}\right), \quad (18)$$

for some constant c.

Assumption (a) is the same as that in Theorem 2, which restricts the noise level. Assumption (b) is about the strength of the tail. A lower β indicates a heavier tail. For example, for one dimensional Gaussian or exponential distribution, $\beta = 1$, while for Cauchy distribution, $\beta = 1/2$. Assumption (c) assumes that η is smooth. Assumption (d) is the minimal mass assumption, which has been used in (Gadat, Klein, and Marteau 2016).

From (18), it can be observed that if

$$k_{max} \gtrsim N^{\min\{\frac{2\beta}{1+\alpha}, \frac{4\beta}{\beta d+2(\alpha+2\beta)}\}},$$

then the first term dominates and hence (18) matches the minimax lower bound derived in (Zhao and Lai 2019) up to a log polynomial factor. Moreover, using the same steps as the proof of Theorem 3, we can also show that our new method is nearly minimax optimal for a slightly different case discussed in (Gadat, Klein, and Marteau 2016), which assumes (a), (b) and (d) in Theorem 3, while (c) is changed to the assumption that η is Lipschitz.

The results in Theorems 2 and 3 and the corresponding discussions show that our new adaptive method is nearly minimax rate optimal under a large variety of settings. From (16) and (18), we also find that it is not always necessary to use large k_{max}. Letting k_{max} grows sublinearly with N instead of linearly with N can help us to reduce the computation cost, without having a significant impact on the classification accuracy. However, considering that the minimum necessary k_{max} depends on the parameters of distributions such as α, β and γ, if these parameters are unknown, it would be better to use a larger k_{max}.

5 Fast Convergence Rate for a Special Class of Distributions

Curse of dimensionality is a common problem for nonparametric learning methods. Minimax analysis in (Audibert and Tsybakov 2007; Gadat, Klein, and Marteau 2016; Zhao and Lai 2019) shows that this problem can not be solved in general. However, we show that for a particular class of distributions, the convergence rate of the new adaptive kNN method does not decay with the increase of dimensionality.

Theorem 4 *Suppose that f and η satisfy the following assumptions: There exist some positive constants $A, c_f, C_f, c_\eta, C_\eta, M_1, M_2, \delta, D$, such that*

(a) For all $\mathbf{x}$ such that $|x_1| < A$, in which x_j is the j-th component of $\mathbf{x}$, $j = 1, \ldots, d$, we have

$$\eta(-x_1, x_2, \ldots, x_d) = -\eta(x_1, \ldots, x_d),$$

and

$$f(-x_1, x_2, \ldots, x_d) = f(x_1, \ldots, x_d);$$

(b) If $|x_1| < A$, then

$$c_f \leq f(x_1 | x_2, \ldots, x_d) \leq C_f;$$

(c) If $|x_1| < A$, then

$$c_\eta \leq \partial\eta(\mathbf{x})/\partial x_1 \leq C_\eta;$$

(d) If $|x_1| < A$, then

$$\|\nabla f\| \leq M_1;$$

(e) If $|x_1| < A$, then

$$\left\|\nabla^2 \eta\right\| \leq M_2;$$

(f) If $|x_1| \geq A/2$, then

$$|\eta(\mathbf{x})| \geq \delta.$$

Let k_{max} grows linearly with N, then
(1) If

$$P(B(\mathbf{x}, r)) \geq f_L r^d$$

for some constant f_L and all $\mathbf{x}$ and $r \leq D$, then

$$R - R^* = \mathcal{O}\left(N^{-1} \ln^2 N\right); \quad (19)$$

(2) If

$$P(B(\mathbf{x}, r)) \geq C_d r^d f(\mathbf{x})$$

for some constant C_d and all $\mathbf{x}$ and $r \leq D$,

$$P(f(\mathbf{X}) \leq t) \leq C_b t^\beta,$$

then

$$R - R^* = \mathcal{O}\left(N^{-\frac{2\beta}{2\beta+1}} \ln^2 N + N^{-\beta} \ln^{2\beta} N\right). \quad (20)$$

Remark 1 *Assumptions (a)-(f) hold for the case in which the underlying distribution is a random mixture of two distributions with opposite label, and these two distributions are the same except that the means are different, i.e.*

$$f(\mathbf{x}|Y = i) = \phi(\mathbf{x} - i\mathbf{c}), i = -1, 1, \quad (21)$$

in which $\mathbf{c}$ is a fixed vector, and

$$P(Y = 1) = P(Y = -1) = 1/2. \quad (22)$$

Remark 2 *In Theorem 4, we assume that η is antisymmetric and f is symmetric around $x_1 = 0$. In fact, the axis $x_1 = 0$ can be generalized to arbitrary linear $(d-1)$ dimensional subspace. In this case, we just need to conduct a simple transformation from $(x_1, \ldots, x_d)$ to $(x'_1, \ldots, x'_d)$, such that the axis becomes $x'_1 = 0$.*

Moreover, assumption (c) can also be changed to

$$c_\eta < -\partial\eta(\mathbf{x})/\partial x_1 \leq C_\eta,$$

then (19) *and* (20) *still hold.*

From (19) and (20), we can observe that, if the pdf is bounded away from zero, then from Theorem 4, the convergence rate is always $\tilde{\mathcal{O}}(N^{-1})$, while for distributions with tails, the convergence rate depends on the tail parameter β. For both cases, it can be observed that the convergence rate of this new adaptive kNN method is faster than those derived in Sections 3 and 4, and does not decrease with the increase of the dimensionality, as long as Assumptions (a)-(f) are satisfied. The complexity of the classification problem under these assumptions is lower than those in Theorem 2 or Theorem 3, since the Bayes decision boundary is linear. As a result, with an adaptive selection of k, the convergence rate can become much faster than the minimax lower bounds derived in (Audibert and Tsybakov 2007; Gadat, Klein, and Marteau 2016; Zhao and Lai 2019).

6 Numerical Examples

In this section, we use numerical experiments to validate our theoretical analysis. In particular, we calculate the convergence rates of the adaptive kNN classifier for some common distributions, and create a log-log plot of the estimated excess risk of the new adaptive kNN classifier versus the sample size. Each point in the curves is averaged over 5,000 trials. The results are shown in Figures 1 and 2.

In Figure 1, we show some examples that satisfy the assumptions in Theorem 2 or 3. In subfigures (a), (b), (c) and (d), the feature distribution is Uniform distribution in $[-5,5]^d$, standard Gaussian distribution, standard Laplace distribution, and triangular distribution in $[-5,5]$ with mode at 0, respectively. From subfigures (a) to (d), the regression functions are all sinusoidal, i.e. $\eta(\mathbf{x}) = \sin(x_1)$. Subfigure (a) is an example that satisfies assumptions in Theorem 2, while subfigures (b), (c) and (d) are examples that satisfy Theorem 3.

In Figure 2, we show some examples that satisfy assumptions in Theorem 4. In particular, in subfigure (a),

$$\eta(\mathbf{x}) = \begin{cases} 2x_1 & \text{if} \quad |x_1| \le \frac{1}{2} \\ 1 & \text{if} \quad x_1 > \frac{1}{2} \\ -1 & \text{if} \quad x_1 < -\frac{1}{2}, \end{cases} \tag{23}$$

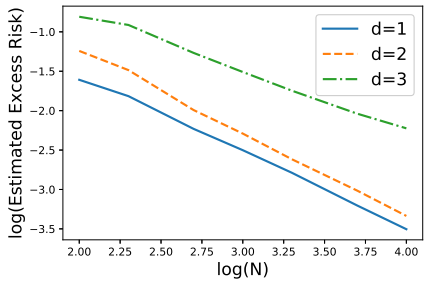
in which x_1 is the first component of $\mathbf{x}$. In subfigures (b), (c) and (d), we analyze the case that the feature distributions for $Y = 1$ and $Y = -1$ are the same except that they have different means:

$$f(\mathbf{x}|Y=1) = \phi(\mathbf{x} - \mathbf{e}_1), \tag{24}$$

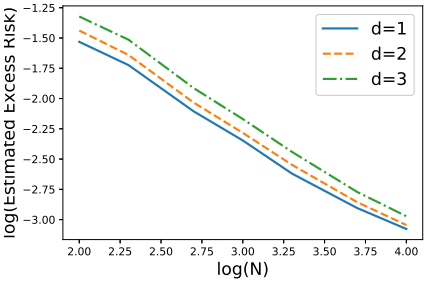
$$f(\mathbf{x}|Y=-1) = \phi(\mathbf{x} + \mathbf{e}_1), \tag{25}$$

in which $\mathbf{e}_1 = (1, 0, \ldots, 0)$ is the unit vector in the first dimension, ϕ can be the pdf of some common distributions, such as Gaussian distribution. This type of distributions satisfy the assumptions in Theorem 4. In subfigures (b), (c) and (d), ϕ is the pdf of the standard Gaussian distribution, standard Laplace distribution and triangular distribution in $[-2,2]$ with mode at 0, respectively.

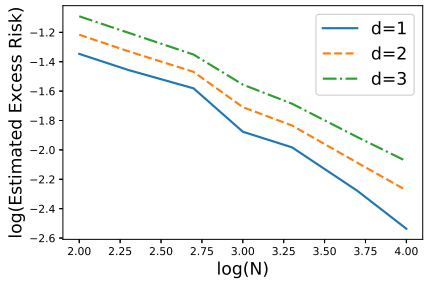
Based on Figures 1 and 2, we calculate the empirical convergence rates of the new adaptive kNN method, which are the negative slopes of the curves in the figures. The empirical rates are then compared with the theoretical rates. The results



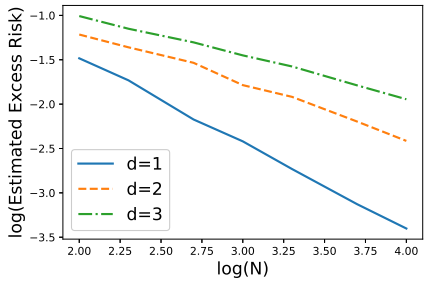
(a) Uniform distributions with sinusoidal regression function.



(b) Gaussian distributions with sinusoidal regression function.

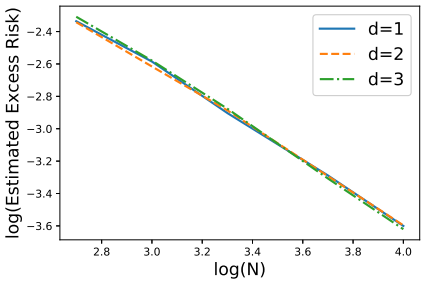


(c) Laplace distributions with sinusoidal regression function.

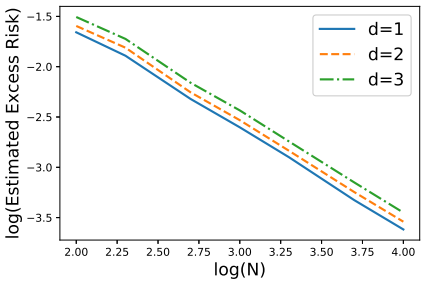


(d) Triangular distributions with sinusoidal regression function.

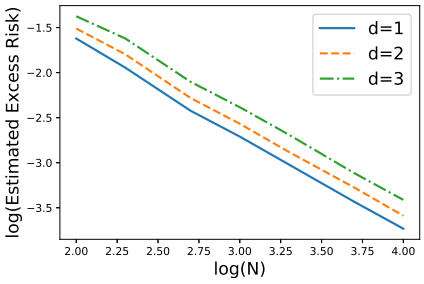
Figure 1: The excess risk vs $\log(N)$ for some examples satisfying assumptions in Theorem 2 or 3.



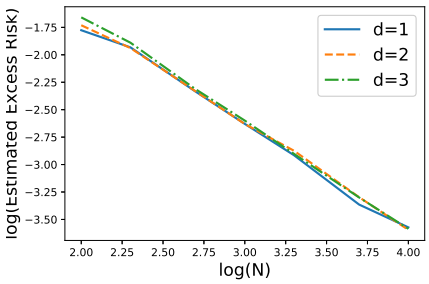
(a) Uniform distribution.



(b) Gaussian distributions with different means.



(c) Laplace distributions with different means.



(d) Triangular distributions with different means.

Figure 2: The excess risk vs $\log(N)$ for some examples satisfying assumptions in Theorem 4.

are shown in Table 1. For the convenience of expression, we use value θ to indicate that the theoretical convergence rate is $\mathcal{O}(N^{-\theta}\ln^c N)$ for some constant c. The theoretical rate of Fig. 1(a) comes from Theorem 2 with $\gamma = 2/d$, while Fig. 1 (b)-(d) comes from Theorem 3 with $\beta = 1, 1, 2$ respectively. For Fig. 2, the theoretical results come from Theorem 4, in which (a) comes from Theorem 4(1), while the remainders come from Theorem 4(2). Note that the margin parameter is $\alpha = 1$ for all of the eight cases, therefore α is not listed in the table.

Table 1: The convergence rate of the new kNN adaptive classifier

Case	Empirical			Theoretical		
	$d=1$	$d=2$	$d=3$	$d=1$	$d=2$	$d=3$
1(a)	0.98	1.03	0.74	0.80	0.67	0.57
1(b)	0.80	0.81	0.83	0.57	0.50	0.44
1(c)	0.60	0.54	0.50	0.57	0.50	0.44
1(d)	0.96	0.60	0.49	0.67	0.57	0.50
2(a)	0.99	0.99	1.00	1.00	1.00	1.00
2(b)	0.99	0.99	0.99	0.67	0.67	0.67
2(c)	1.05	1.04	1.03	0.67	0.67	0.67
2(d)	0.97	0.97	0.98	0.80	0.80	0.80

From Table 1, we can observe that for some cases, the empirical rates are approximately the same as the theoretical bounds, while for most of the other cases, the empirical rates are actually faster. As discussed in Section 4, under the assumptions in Theorem 2 and 3, the convergence rates are already minimax optimal. However, the minimax lower bounds in (Audibert and Tsybakov 2007; Zhao and Lai 2019) are established for the worst case that satisfies the assumptions. Our numerical results show that for many common distributions, the real convergence rates can be much faster. In particular, if the assumptions (a)-(f) in Theorem 4 hold, then the convergence rates are fast and do not decay with the increase of dimensionality. These assumptions hold commonly for cases that are constructed by random mixtures of two distributions with opposite labels, different means but the same shapes.

7 Conclusion

In this paper, we have proposed a new adaptive kNN classifier, which selects different k for different test samples. Our analysis has shown that it is minimax optimal up to a log polynomial factor under some assumptions. Moreover, if the Bayes decision boundary is linear, under some other assumptions, the convergence rate can be faster, and does not become slower with the increase of the dimensionality. Compared with previous adaptive kNN methods, this method is more convenient to use since it does not require any knowledge of the underlying distribution. The performance of our new method is also competitive, since for many common distributions, the real convergence rate is faster than the minimax lower bound. Numerical experiments have been conducted to validate our theoretical analysis.

References

Audibert, J.-Y. 2004. Classification under polynomial entropy and margin assumptions and randomized estimators Preprint.

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

Balsubramani, A.; Dasgupta, S.; Moran, S.; et al. 2019. An adaptive nearest neighbor rule for classification. In *Advances in Neural Information Processing Systems*, 7577–7586.

Bentley, J. L.; Stanat, D. F.; and Williams Jr, E. H. 1977. The complexity of finding fixed-radius near neighbors. *Information processing letters* 6(6): 209–212.

Biau, G.; and Devroye, L. 2015. *Lectures on the nearest neighbor method*. Springer.

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

Cérou, F.; and Guyader, A. 2006. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics* 10: 340–355.

Chaudhuri, K.; and Dasgupta, S. 2014. Rates of convergence for nearest neighbor classification. In *Proc. Advances in Neural Information Processing Systems*, 3437–3445. Montreal, Canada.

Cover, T.; and Hart, P. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1): 21–27.

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

Devroye, L.; Györfi, L.; and Lugosi, G. 2013. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Döring, M.; Györfi, L.; and Walk, H. 2017. Rate of convergence of k-nearest-neighbor classification rule. *Journal of Machine Learning Research* 18(1): 8485–8500.

Fix, E. 1951. *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine.

Gadat, S.; Klein, T.; and Marteau, C. 2016. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *The Annals of Statistics* 44(3): 982–1009.

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

Kohler, M.; and Krzyzak, A. 2007. On the rate of convergence of local averaging plug-in classification rules under a margin condition. *IEEE Transactions on Information Theory* 53(5): 1735–1742.

Kpotufe, S. 2011. k-NN regression adapts to local intrinsic dimension. In *Proc. Advances in Neural Information Processing Systems*, 729–737. Granada, Spain.

Leibe, B.; Mikolajczyk, K.; and Schiele, B. 2006. Efficient clustering and matching for object class recognition. In *BMVC*, 789–798.

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

Ougiaroglou, S.; Nanopoulos, A.; Papadopoulos, A. N.; Manolopoulos, Y.; and Welzer-Druzovec, T. 2007. Adaptive k-nearest-neighbor classification using a dynamic number of nearest neighbors. In *East European Conference on Advances in Databases and Information Systems*, 66–82. Springer.

Stone, C. J. 1977. Consistent nonparametric regression. *The Annals of Statistics* 595–620.

Sun, S.; and Huang, R. 2010. An adaptive k-nearest neighbor algorithm. In *2010 seventh international conference on fuzzy systems and knowledge discovery*, volume 1, 91–94. IEEE.

Zhao, P.; and Lai, L. 2019. Minimax Rate Optimal Adaptive Nearest Neighbor Classification and Regression. *arXiv preprint arXiv:1910.10513* .