k-Nearest Neighbour Classification of Datasets with a Family of Distances

Stan Hatko

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

The k-nearest neighbour (k-NN) classifier is one of the oldest and most important supervised learning algorithms for classifying datasets. Traditionally the Euclidean norm is used as the distance for the k-NN classifier. In this thesis we investigate the use of alternative distances for the k-NN classifier.

We start by introducing some background notions in statistical machine learning. We define the k-NN classifier and discuss Stone's theorem and the proof that k-NN is universally consistent on the normed space $(\mathbb{R}^d, \|\cdot\|)$. We then prove that k-NN is universally consistent if we take a sequence of random norms (that are independent of the sample and the query) from a family of norms that satisfies a particular boundedness condition. We extend this result by replacing norms with distances based on uniformly locally Lipschitz functions that satisfy certain conditions. We discuss the limitations of Stone's lemma and Stone's theorem, particularly with respect to quasi-norms and adaptively choosing a distance for k-NN based on the labelled sample. We show the universal consistency of a two stage k-NN type classifier where we select the distance adaptively based on a split labelled sample and the query. We conclude by giving some examples of improvements of the accuracy of classifying various datasets using the above techniques.

Acknowledgements

I would like to thank my supervisor Dr. Vladimir Pestov for his guidance and support during my studies and for introducing me to the field of machine learning.

I would also like to thank the members of the University of Ottawa Data Science Group and people who worked with the group, including Gaël Giordano, Hubert Duan, Yue Dong, Samuel Buteau, Julien Roy, Sabrina Sixta, Émilie Idene, Luiz Gustavo Cordeiro, and Christian Despres.

Finally, I would like to thank my family for their support.

Dedication

To my parents.

Contents

List of Figures

List of Tables

Chapter 1

An Introduction to Statistical Machine Learning

In this chapter we introduce the fundamental notions of statistical machine learning. No prior knowledge of statistical machine learning is assumed in this section. We start by giving an informal discussion with some examples and then discussing the theory on a more formal level.

1.1 An Informal Introduction

In the classification problem of statistical machine learning, we start with a dataset (where the points come from some sample space), together with a label (or class) for each point (where there are a finite number of possible labels). We suppose that the points in the dataset are independently and identically distributed. We have a new data point, called the query, from the same distribution as the data set, and which is also assumed to be independent of the points in the data set. However, we do not have the label for the query. We would like to predict the label for the query based on the dataset.

For instance, suppose we would like to predict if a person has a predisposition for heart disease based on their genome. We have a dataset of the genome of people with their genomic sequence and whether or not they have heart disease. We now have a new patient, for which we have the genome but do not know if they have heart disease. We would like to predict, based on their genomic sequence, if they have heart disease, with the only information available to us being the dataset and the person's genomic sequence.

Let X be the dataset and Y be a set of classes. A *classifier* $f : X \to Y$ is a function that attempts to predict a class y for a data point x. The *accuracy* of the classifier f is the probability that we will predict the correct label for the query, and the *misclassification error* of f is the probability that we will predict a wrong label. Given the query point, we would like to predict its label. We would like to find a classifier f whose accuracy is as high as possible (or equivalently, whose error is as small as possible). The *Bayes error* is the infimum of the errors of all possible classifiers for a distribution μ on $X \times Y$. We can show that the Bayes error is attained by the *Bayes classifier*, however constructing the Bayes classifier requires knowledge of the underlying distribution μ, which we normally do not have, we only have a set of labelled data points.

The process of constructing a classifier f is called *learning*. A *learning rule* is a family of functions that takes a set of labelled data points and outputs a classifier, which we can then use to classify query points. A learning rule is said to be *consistent* for a distribution μ if the expected value of the error converges to the Bayes error in probability for μ as the number of labelled data points goes to infinity. A learning rule is *universally consistent* if it is consistent for every distribution μ on $X \times Y$. Common learning rules include those based on k-nearest neighbour, support vector machine (SVM), and random forest. When applying a learning rule and then using it to classify points, we often refer to the combination of the learning rule and classifier together as simply a classifier.

For any classifier, to test its accuracy we take the dataset and split it into two disjoint subsets, the *training set* and the *testing set*. The training set is used in constructing f, and from this we predict the labels for points in the testing set. We then compare the predicted labels to the correct labels in the testing set and compute the accuracy of our prediction.

1.2 Theory of Statistical Machine Learning

We now introduce the fundamental notions of the theory of statistical machine learning. Let Ω be a nonempty set called the *domain*, $\{1, 2, \ldots, q\}$ (with $q \geq 2$) be a finite set of *labels* (or *classes*), and μ be a probability measure on $\Omega \times \{1, 2, \ldots, q\}$. We often assume without loss of generality that there are only two classes (the *binary classification problem*), for this section we will consider the case of $q \geq 2$ classes, but afterwards we will focus on the $q = 2$ case.

A *classifier* is a Borel measurable function $f : \Omega \to \{1, 2, \ldots, q\}$, that maps points in the domain Ω to classes in $\{1, 2, \ldots, q\}$. We define the *misclassification error* of a classifier as the probability that the label predicted by our classifier is different than the true label,

$$\mathrm{err}_\mu(f) = \mu\left(\{(x, y) \in \Omega \times \{1, 2, \ldots, q\} \ : \ f(x) \neq y\}\right). \tag{1.1}$$

The *Bayes error* is the infimum of the misclassification error over all possible classifiers for the probability measure μ on $\Omega \times \{1, 2, \ldots, q\}$,

$$\ell^*(\mu) = \inf_f \mathrm{err}_\mu(f). \tag{1.2}$$

We see that since the misclassification error must be in $[0, 1]$ (since any probability must be in $[0, 1]$), and the set of classifiers is nonempty (since we can simply take the classifier that maps every point in Ω to zero), it follows that the infimum exists in $[0, 1]$ and so the Bayes error is always well defined.

Suppose we have a set of n independent and identically distributed random ordered pairs $D_n = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, modelling the data. A *learning rule* $\mathcal{L} = (\mathcal{L}_n)_{n=1}^{\infty}$ is a family of functions that maps each possible labelled sample to a classifier,

$$\mathcal{L}_n : (\Omega \times \{1, 2, \ldots, q\})^n \to \{f : \Omega \to \{1, 2, \ldots, q\} \mid f \text{ is Borel}\}. \tag{1.3}$$

Common learning rules include those based on k-nearest neighbour (k-NN), Support Vector Machine (SVM), and Random Forest. In applications when classifying datasets (so we classify points immediately when learning) it is common to simply refer to these as "classifiers", for now we will continue to make the distinction between learning rules and classifiers. A learning rule can also be thought of as a sequence of classifiers constructed based on the labelled sample of points.

Let $D_n = (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be an iid labelled sample with distribution μ, and (X, Y) be the query and the label, which is independent from the sample and also has the same distribution μ. We let $\ell^*(\mu)$ be the Bayes error for the distribution μ. Given the labelled sample, the error probability is the conditional probability

$$L_n = \mathbb{P}(\mathcal{L}_n(X, D_n) \neq Y | D_n). \tag{1.4}$$

A learning rule $\mathcal{L}$ is said to be *consistent* (or *weakly consistent*) for the distribution μ if the misclassification error of the learning rule $(\mathcal{L}_n)_{n=1}^{\infty}$ converges in probability to the Bayes error, that is, as $n \to \infty$,

$$\lim_{n \to \infty} \mathbb{P}(|L_n - \ell^*(\mu)| > \epsilon) = 0 \tag{1.5}$$

or equivalently, that

$$\mathbb{P}\left(\mathcal{L}_n(X, D_n) \neq Y\right) \to \ell^*(\mu). \tag{1.6}$$

We say that $\mathcal{L}$ is *strongly consistent* if with probability one we have a sequence of labelled samples $D_1, D_2, \ldots$ such that the misclassification error approaches the

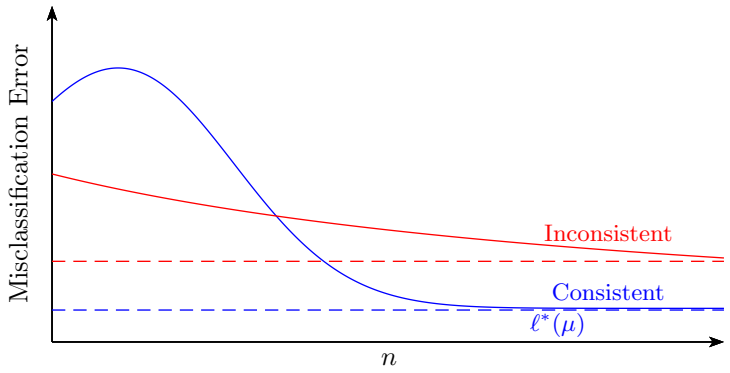

Figure 1.1: Illustration of the misclassification error of a consistent and an inconsistent learning rule. We see that the misclassification error of the consistent learning rule approaches the Bayes error $\ell^*(\mu)$ as the samples size n approaches infinity, while the inconsistent rule (which in this case starts off better for small n) performs more poorly as n increases and does not converge to the Bayes error.

Bayes error as the sample size n approaches infinity (so the above convergence in probability is replaced by almost sure convergence). That is, we have

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbb{P}(\mathcal{L}_n(X, D_n) \neq Y | D_n) = \ell^*(\mu)\right) = \mathbb{P}\left(\lim_{n\to\infty} L_n = \ell^*(\mu)\right) = 1. \tag{1.7}$$

If the learning rule is consistent for every distribution on $\Omega \times \{1, 2, \ldots, q\}$, we say that it is *universally consistent*. We define *strong universal consistency* in the same way, that a learning rule is strongly consistent for every distribution on $\Omega \times \{1, 2, \ldots, q\}$. [1]

A learning rule whose misclassification error is monotone decreasing at each step n is called a *smart* learning rule. A simple example of a learning rule that is "smart" by this definition is to select the classifier that selects a particular fixed label always (completely ignoring the labelled sample), then the misclassification error is constant

[1] Not every learning rule used in applications is universally consistent, for instance, random forests are not universally consistent ([36], Proposition 8) but have a very good classification accuracy on many datasets and are commonly used in applications.

regardless of the sample size (this is not a learning rule we would call "smart" in the usual sense of the word). Such a learning rule is obviously not universally consistent. A universally consistent learning rule is not necessarily smart, the misclassification error can temporarily increase for some n before decreasing again towards the Bayes error. There are no known examples of a universally consistent smart learning rule, it has been conjectured that no such learning rules exist.

1.2.1 The Regression Function and the Bayes Classifier

In this section, we assume we have a binary classification problem, that is, the set of classes is $\{0, 1\}$. We let μ be a probability measure on $\Omega \times \{0, 1\}$. We then define two new measures ν, λ on Ω by (for any Borel set $A \subseteq \Omega$):

$$\nu(A) = \mu(A \times \{1\}) \tag{1.8}$$

$$\lambda(A) = \mu(A \times \{0, 1\}) = \mu(A \times \{0\}) + \mu(A \times \{1\}) \tag{1.9}$$

We observe that for any Borel set $A \subseteq \Omega$, $\nu(A) \leq \lambda(A)$, and so ν is absolutely continuous with respect to λ. Hence by the Radon-Nikodym derivative theorem, the Radon-Nikodym derivative of ν with respect to λ exists, which we call the regression function η (that is, for any Borel set A, $\int_A \eta \mathrm{d}\lambda = \nu(A)$).[27, 14] By the Radon-Nikodym derivative theorem, η is integrable with respect to λ and is Borel measurable. Equivalently, we can also write η as the conditional probability $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

With the regression function, we are now able to define a classifier called the *Bayes classifier*. The Bayes classifier g^* is defined as:

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{1.10}$$

We now show that the Bayes classifier is optimal, that is, it has the highest

accuracy of any classifier on our dataset. This is a standard result, the proof below is based on the one found in [14] (Theorem 2.1) and [18] (Theorem 1.1.2).

Theorem 1.2.1 (Bayes Optimality Theorem)**.** *For any classifier $g : \Omega \to \{0, 1\}$ and any probability distribution μ on $\Omega \times \{0, 1\}$, we have the inequality*

$$\mu(\{(x, y) \ : \ g^*(x) \neq y\}) \leq \mu(\{(x, y) \ : \ g(x) \neq y\}). \tag{1.11}$$

Equivalently, we can write this in terms of random variables,

$$\mu(g^*(X) \neq Y) \leq \mu(g(X) \neq Y). \tag{1.12}$$

From this, we see that the expected error of the Bayes classifier is the infimum of the misclassification errors of any classifier (for the distribution μ on $\Omega \times \{0, 1\}$). Hence the Bayes classifier achieves the Bayes error, and any classifier has a misclassification error which is at least that of the Bayes classifier.

Proof: It suffices for us to show that for all $x \in \Omega$,

$$\mu(g^*(X) \neq Y | X = x) \leq \mu(g(X) \neq Y | X = x). \tag{1.13}$$

For any classifier $g : \Omega \to \{0, 1\}$, the following holds:

$$\begin{aligned}
\mu(g(X) \neq Y | X = x) &= 1 - (\mu(Y = 1, g(x) = 1 | X = x) + \mu(Y = 0, g(x) = 0 | X = x)) \\
&= \begin{cases} 1 - \mu(Y = 1 | X = x) & \text{if } g(x) = 1 \\ 1 - \mu(Y = 0 | X = x) & \text{if } g(x) = 0 \end{cases} \\
&= 1 - \eta(x)^{g(x)}(1 - \eta(x))^{1 - g(x)}
\end{aligned}$$

We see that the above equality holds with g^* as well, so we have:

$$\begin{aligned}
&\mu(g(X) \neq Y | X = x) - \mu(g^*(X) \neq Y | X = x) \\
&= 1 - \eta(x)^{g(x)}(1 - \eta(x))^{1 - g(x)} - \left(1 - \eta(x)^{g^*(x)}(1 - \eta(x))^{1 - g^*(x)}\right)
\end{aligned}$$

$$= \eta(x)^{g^*(x)}(1-\eta(x))^{1-g^*(x)} - \eta(x)^{g(x)}(1-\eta(x))^{1-g(x)}$$

$$= \begin{cases} 0 & \text{if } g(x) = g^*(x) \\ 2\eta(x) - 1 & \text{if } g^*(x) = 1 \text{ and } g(x) = 0 \\ 1 - 2\eta(x) & \text{if } g^*(x) = 0 \text{ and } g(x) = 1 \end{cases}$$

Since $g^*(x) = 1$ if and only if $\eta(x) \geq 1/2$, we have:

- $2\eta(x) - 1 > 0$ when $g^*(x) = 1$.
- $1 - 2\eta(x) \geq 0$ when $g^*(x) = 0$.

We therefore have

$$\mu(g(X) \neq Y | X = x) - \mu(g^*(X) \neq Y | X = x) \geq 0.$$

Hence we have that equation (1.13) holds, and so the theorem is proven. ∎

In order to construct the regression function η, we need to know the underlying distribution μ, which we do not have access to. This means we cannot compute the regression function η directly and simply use the Bayes classifier. The regression function is still a powerful theoretical notion which is very useful in proving various inequalities. We will often consider various estimates to the regression function, some of which can be constructed empirically from the data set.

1.2.2 An Example

We now illustrate a classical and simple example of a learning rule and classifier, and show it is consistent for a distribution but is inconsistent for another distribution. Suppose we have the distribution μ on $[0,1] \times \{0,1\}$ that takes $(0,0)$ with probability $1/2$ (that is, a point mass at zero, with label zero), and otherwise (with probability $1/2$) is uniformly distributed on $(0,1]$ with label 1. That is, there is a point mass at

0 with label 0, and otherwise it is uniformly distributed on the rest of the interval with label 1. We see that the regression function $\eta : [0,1] \to [0,1]$ is

$$\eta(x) = \begin{cases} 0 & \text{if } x = 0 \\ 1 & \text{otherwise.} \end{cases} \tag{1.14}$$

Hence, the Bayes classifier classifies the point 0 as having label 0, and any other point as label 1. The Bayes error is zero, since the label is a deterministic function of the point.

A simple learning rule is the *nearest neighbour* learning rule (1-NN). In 1-NN, for a query X we assign the label of the nearest point in the dataset to X (we use the usual metric $d(x,y) = |x-y|$ here). For our distribution μ, a point is misclassified if it is at the point $x = 0$ and is assigned label 1 or is not at zero ($x \neq 0$) but is assigned label 0. Suppose we have an iid labelled sample of n points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. There is a $1/2$ probability of the query being at 0 and having label 0. In this case, the query will be misclassified if and only if none of the points in the sample are at zero (that is, all of the points are not at zero and have label 1). The probability of this occurring for a sample of n points is 2^{-n}, which goes to zero as $n \to \infty$. In the other case, with $1/2$ probability, the query is nonzero and has label 1. The only way that the query will be misclassified in this case is if either there are no points in the sample with label 1 or the nearest point with label 1 is further from the query than zero. We see that the probability of either of these occurring goes to zero as n approaches infinity. Hence we find that the point is classified correctly with probability approaching one as $n \to \infty$ (equivalently, the error goes to zero as $n \to \infty$), and so 1-NN is consistent for this distribution. This is an example of learning a *deterministic concept* (that is, the Bayes error is zero), for which 1-NN is always consistent (this is an immediate consequence of Theorem 5.4 in [14]).

Now, suppose we take the distribution ν on $[0,1] \times \{0,1\}$ that is uniform on $[0,1]$ such that the label 0 occurs with probability $1/3$ and label 1 occurs with probability

2/3, with the label being independent of the point in $[0, 1]$ (this is an example of a "probabilistic" or "fuzzy" concept, as opposed to a deterministic concept). We notice that the Bayes error is 1/3 and is attained by predicting the label 1 always. We now find the expected error of the 1-NN classifier. Given a query X, there is a 1/3 chance of the label being zero and a 2/3 chance of the label being one. The nearest neighbour $X_{(1)}$ also has a 1/3 probability of being label zero and a 2/3 probability of being label one, independent of the label of X. In the 1-NN classifier, we assign the label of $X_{(1)}$ to the query X. Hence the probability that we will misclassify X (for any sample size $n \geq 1$) is $(1/3)(2/3) + (2/3)(1/3) = 4/9$, which is greater than 1/3. Hence the 1-NN learning rule is not consistent for the distribution ν. This implies that the 1-NN learning rule is not universally consistent, even though it is consistent for the distribution μ above.

Chapter 2

The k-Nearest Neighbour Classifier

In this chapter we discuss one of the most important learning rules for classifying points, the k-nearest neighbour classifier (k-NN). We first start by briefly discussing k-NN with an example, we then give a precise mathematical formulation of k-NN and we present the proof that it is universally consistent (provided that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$).

2.1 The k-Nearest Neighbour Classifier

Suppose we have a set of points in a metric space Ω, with each point assigned a label 0 or 1. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a labelled sample and let (X, Y) be the query. In the k-nearest neighbour classifier, we predict the label of the query based on which class is more common among the k closest points to X in the labelled sample. We illustrate an example of this in Figure 2.1.

We have selected k to be odd in our example to avoid the case of ties. There are two possible cases where ties can occur in our algorithm: it is possible to have multiple classes occurring equally frequently among the k-nearest neighbours of the query, and it is possible to have distances ties with multiple points at the same distance from

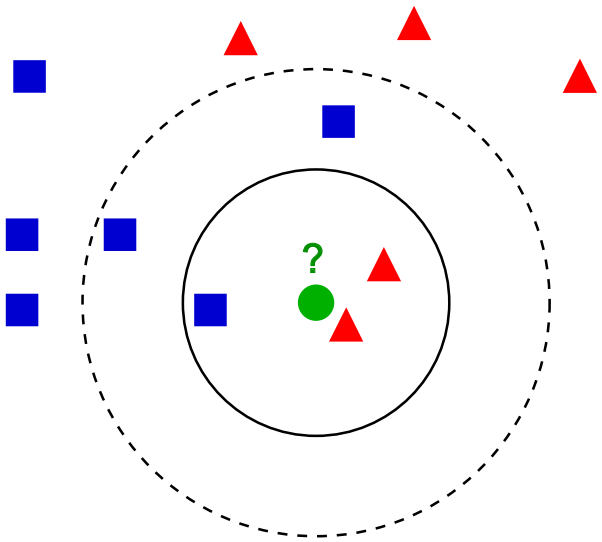

Figure 2.1: An example of k-NN. We classify the query as a triangle for $k = 3$, and as a square for $k = 5$. Image from [23].

the query. Many authors discuss consistency for distributions with a density to avoid the case of distance ties, however, we will prove universal consistency here and will not make such assumptions. Various methods for breaking ties are discussed in the literature. One common way to break ties is by random selection, so that if a voting tie occurs we pick randomly from the most common labels, and if a distance tie occurs we pick a random point at that distance. For our purposes for the binary classification problem, we will break voting ties (with the same number of points in each class, for a given k) by simply selecting the label 1. We break distance ties by generating independent random variables $U_1, U_2, \ldots, U_n$ from the uniform distribution on $[0, 1]$, if there is a distance tie between two points X_i and X_j, we select X_i if $U_i > U_j$ and X_j if $U_j > U_i$ (we can ignore ties between U_i and U_j, since the probability that $U_i = U_j$ is zero). Example pseudocode of k-NN is shown in Algorithm 1.

We would like to establish that the k-NN classifier is universally consistent with the data points being independent and identically distributed. There are at least two known ways to do this, the first is the original proof by Stone which uses Stone's

Algorithm 1 k-NN pseudocode

Require: $k \in \mathbb{N}$, X is the domain, Y is the response (must be a finite set $\{1, 2, ..., p\}$), $a \in X$, (x_1, y_1), ..., $(x_n, y_n) \in X \times Y$

{Calculate distances from input point to all the data points}

for $i = 1$ **to** n **do**

$d_i \leftarrow d(a, x_i)$

end for

{Find response for the k nearest neighbours of the input point}

for $i = 1$ **to** k **do**

$m \leftarrow \arg\min_m \{d_m \text{ such that } 1 \leq m \leq n \text{ not previously selected}\}$

$a_i \leftarrow y_m$

end for

{Find number of times each response occurs among the k nearest neighbours}

for $i = 1$ **to** p **do**

$v_i \leftarrow$ number of times i occurs in $\{a_1, a_2, ..., a_p\}$

end for

$r \leftarrow \{y_i | 1 \leq i \leq p \text{ such that } v_i \text{ is maximal among } v_1, v_2, ..., v_p\}$ {Find the most common response among the k nearest neighbours, if multiple responses are the most common, pick a fixed one}

return r {Return most common response (or if a tie occurs, one of the most common responses)}

theorem, which we state and prove below, and another is the alternative proof that uses the Lebesgue-Besicovitch differentiation theorem, which was originally done in [34] and further discussed in [24].

2.2 Stone's Theorem

The original way in which k-NN was shown to be universally consistent was *Stone's theorem*, named after Charles Stone.[2] This was the first time any learning rule was shown to be universally consistent. We show that any classifier of a particular form that satisfies certain conditions is universally consistent, and then show that the k-NN classifier satisfies these conditions. We prove a slightly stronger version of the original Stone's theorem (the slight strengthening will be used later to assist in the proof of some results). Stone's theorem is the foundation for the results we will prove later on, that is why we discuss the proof (of Stone's theorem and the universal consistency of k-NN) in detail (following the approach in [14] and [18]).

Let Ω be the domain, with μ being a probability measure and η be the regression function on $\Omega \times \{0,1\}$. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a labelled sample. We define real-valued *weights* $W_{ni}(X, (X_1, Y_1), \ldots, (X_n, Y_n), U_1, \ldots, U_n, V)$ that are functions of the query X, the labelled sample, the tiebreakers $U_1, \ldots, U_n$, and possibly a random variable V that is independent of all the other random variables, such that they are nonnegative and sum to one,

$$\sum_{i=1}^{n} W_{ni}(x) = 1. \tag{2.1}$$

We then define the estimate η_n to the true regression function η as the sum of the positive entries multiplied by their weights,

$$\eta_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x). \tag{2.2}$$

We now define a classifier $\mathcal{L}_n$ as follows:

$$\mathcal{L}_n(x) = \begin{cases} 1 & \text{if } \eta_n(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Given a query X, we define $X_{(1)}, \ldots, X_{(n)}$ to be the points $X_1, \ldots, X_n$ in order of increasing distance from X (in the case of a tie between distances, we generate independent uniform random variables $U_1, U_2, \ldots, U_n$ on $[0, 1]$ and we take the point X_i such that the corresponding U_i is larger). In the k nearest neighbour (k-NN) learning rule, we take the weights $W_{ni}(X)$ to be $1/k$ if $X_i \in \{X_{(1)}, \ldots, X_{(k)}\}$ and 0 otherwise.

We now prove a couple of inequalities, which will be useful for us.

Lemma 2.2.1. *For all $a, b, c \in \mathbb{R}$, we have the inequalities*

1. $(a+b)^2 \leq 2(a^2+b^2)$.
2. $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$.

Lemma 2.2.2. *The expected value of the difference of the value of the regression function at X_i and Y_i is zero,*

$$\mathbb{E}\left[\eta(X_i) - Y_i\right] = 0. \tag{2.4}$$

Proof: We see that (since Y_i is nonzero if and only if $Y_i = 1$)

$$\begin{aligned} \mathbb{E}\left[\eta(X_i) - Y_i\right] &= \mathbb{E}\left[\eta(X_i)\right] - \mathbb{E}\left[Y_i\right] \\ &= \mathbb{E}\left[\mathbb{P}(Y_i = 1 | X_i)\right] - \mathbb{E}\left[Y_i\right] \\ &= \mathbb{E}\left[\mathbb{E}[Y_i | X_i]\right] - \mathbb{E}\left[Y_i\right] \\ &= \mathbb{E}\left[Y_i\right] - \mathbb{E}\left[Y_i\right] \\ &= 0. \end{aligned}$$

∎

We now have a lemma that lets us bound the difference of the expected error and the Bayes error. This is a standard result, the proof below is based on [14] (Theorem 6.5) and [18] (Theorem 2.2.5), with more details explained.

Lemma 2.2.3. *If a classifier $\mathcal{L}_n$ is defined as in equation* (2.3)*, then the error probability satisfies the inequalities*

$$\operatorname{err}(\mathcal{L}_n) - \ell^* \leq 2\mathbb{E}\left[|\eta(X) - \eta_n(X)|\right] \tag{2.5}$$

and

$$\operatorname{err}(\mathcal{L}_n) - \ell^* \leq 2\sqrt{\mathbb{E}\left[(\eta(X) - \eta_n(X))^2\right]}. \tag{2.6}$$

Proof: From our proof of Theorem 1.2.1, we have (where $\mathcal{L}^*$ is the Bayes classifier):

$$\begin{aligned}
&\mathbb{P}(\mathcal{L}_n(X) \neq Y | X = x) - \mathbb{P}(\mathcal{L}^*(X) \neq Y | X = x) \\
&= \begin{cases} 0 & \text{if } \mathcal{L}_n(x) = \mathcal{L}^*(x) \\ 2\eta(x) - 1 & \text{if } \mathcal{L}^*(x) = 1 \text{ and } \mathcal{L}_n(x) = 0 \\ 1 - 2\eta(x) & \text{if } \mathcal{L}^*(x) = 0 \text{ and } \mathcal{L}_n(x) = 1 \end{cases} \\
&= |2\eta(x) - 1| \, \mathbb{1}_{\{\mathcal{L}_n(x) \neq \mathcal{L}^*(x)\}} \quad \begin{array}{l}\text{since the above values are always nonnegative by} \\ \text{the definition of } \mathcal{L}^*\end{array}
\end{aligned}$$

Hence we have

$$\begin{aligned}
\mathbb{P}(\mathcal{L}_n(x) \neq Y) - \ell^*_\mu &= \mathbb{P}(\mathcal{L}_n(X) \neq Y) - \mathbb{P}(\mathcal{L}^*(X) \neq Y) \\
&= \mathbb{E}\left[\mathbb{P}(\mathcal{L}_n(X) \neq Y | X = x)\right] - \mathbb{E}\left[\mathbb{P}(\mathcal{L}^*(X) \neq Y | X = x)\right] \\
&= \mathbb{E}\left[\mathbb{P}(\mathcal{L}_n(X) \neq Y | X = x) - \mathbb{P}(\mathcal{L}^*(X) \neq Y | X = x)\right] \\
&= \int_\Omega |2\eta(x) - 1| \, \mathbb{1}_{\{\mathcal{L}_n(x) \neq \mathcal{L}^*(x)\}} \mathrm{d}\mu(x \times \{0, 1\}) \\
&= 2\int_\Omega |\eta(x) - 1/2| \, \mathbb{1}_{\{\mathcal{L}_n(x) \neq \mathcal{L}^*(x)\}} \mathrm{d}\mu(x \times \{0, 1\}).
\end{aligned}$$

We see that the function we are integrating can only be nonzero when $\mathcal{L}_n(\omega) \neq \mathcal{L}^*(\omega)$. If $\mathcal{L}_n(x) = 1$ and $\mathcal{L}^*(x) = 0$, then $\eta(x) < 1/2$ and $\eta_n(x) \geq 1/2$. Similarly, we

find that if $\mathcal{L}_n(x) = 0$ and $\mathcal{L}^*(x) = 1$ then $\eta(x) \geq 1/2$ and $\eta_n(x) < 1/2$. In both cases we have the inequality

$$|\eta(x) - 1/2| \leq |\eta(x) - \eta_n(x)|\,.$$

Combining the above results, we find

$$\begin{aligned}
\mathbb{P}(\mathcal{L}_n(x) \neq Y) - \ell_\mu^* &= 2\int_\Omega |\eta(x) - 1/2|\, \mathbb{1}_{\{\mathcal{L}_n(x) \neq \mathcal{L}^*(x)\}} \mathrm{d}\mu(x \times \{0,1\}) \\
&\leq 2\int_\Omega |\eta(x) - \eta_n(x)|\, \mathbb{1}_{\{\mathcal{L}_n(x) \neq \mathcal{L}^*(x)\}} \mathrm{d}\mu(x \times \{0,1\}) \\
&\leq 2\int_\Omega |\eta(x) - \eta_n(x)|\, \mathrm{d}\mu(x \times \{0,1\}) \\
&= 2\mathbb{E}\left[|\eta(X) - \eta_n(X)|\right].
\end{aligned}$$

We have now proven the first inequality (2.5). The second inequality (2.6) follows by applying Jensen's inequality, so we find

$$\begin{aligned}
\mathbb{P}(\mathcal{L}_n(x) \neq Y) - \ell_\mu^* &\leq 2\mathbb{E}\left[|\eta(X) - \eta_n(X)|\right] \\
&\leq 2\sqrt{\mathbb{E}\left[(\eta(X) - \eta_n(X))^2\right]}.
\end{aligned}$$

∎

A core result is *Stone's Theorem*, which gives sufficient conditions for $\mathcal{L}_n$ to be universally consistent. We state a slightly strengthened version of Stone's theorem below, the only difference from the original version is that the original does not include the ϵ_n sequence in the first condition and we only require bounded functions in the first condition. The proof of this result is based on [14] (Theorem 6.3) and [18] (Theorem 2.2.2), with more details added.

Theorem 2.2.4 (Stone's Theorem)**.** *Suppose a learning rule $(\mathcal{L}_n)_{n=1}^\infty$ is defined as in* (2.3)*, with the domain Ω being $\mathbb{R}^d$. Then if the following conditions hold (for any probability distribution of $(X,Y), (X_1,Y_1), \ldots, (X_n,Y_n)$ on $\mathbb{R}^d \times \{0,1\}$, with the points being iid), $(\mathcal{L}_n)_{n=1}^\infty$ is universally consistent.*

1. *There exists a constant $c \in \mathbb{R}$ and a sequence $(\epsilon_n)_{n=1}^{\infty}$ that goes to zero, $\epsilon_n \to 0$ as $n \to \infty$, such that for every measurable nonnegative function $f : \mathbb{R}^d \to \mathbb{R}$ bounded above by one, for all $n \geq 1$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}\left[f(X)\right] + \epsilon_n. \tag{2.7}$$

2. *There exists a norm $\|\cdot\|$ such that for all $a > 0$, as $n \to \infty$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > a\}}\right] \to 0. \tag{2.8}$$

3. *As $n \to \infty$,*

$$\mathbb{E}\left[\max_{1 \leq i \leq n} W_{ni}(X)\right] \to 0. \tag{2.9}$$

Proof: For our proof, we show that

$$\mathbb{E}\left[(\eta(X) - \eta_n(X))^2\right] \to 0 \text{ as } n \to \infty. \tag{2.10}$$

By Lemma 2.2.3, this implies that $\operatorname{err} \mathcal{L}_n - \ell^* \to 0$ as $n \to \infty$, and hence that $(\mathcal{L}_n)_{n=1}^{\infty}$ is universally consistent.

- We define another approximation $\hat{\eta}_n$ of the regression function η by

$$\hat{\eta}_n(x) = \sum_{i=1}^{n} \eta(X_i) W_{ni}(x). \tag{2.11}$$

We now see that (by Lemma 2.2.1):

$$\begin{aligned} \mathbb{E}\left[(\eta(X) - \eta_n(X)^2\right] &\leq \mathbb{E}\left[(\eta(X) - \hat{\eta}_n(X) + \hat{\eta}_n(X) - \eta_n(X))^2\right] \\ &\leq 2\mathbb{E}\left[(\eta(X) - \hat{\eta}_n(X))^2\right] + 2\mathbb{E}\left[(\hat{\eta}_n(X) - \eta_n(X))^2\right] \end{aligned} \tag{2.12}$$

We now show that both terms go to zero as $n \to \infty$, by showing that both terms (in (2.12)) can be made arbitrarily small.

- We show that for any $\epsilon > 0$, for sufficiently large n, $\mathbb{E}\left[(\eta(X) - \hat{\eta}_n(X))^2\right] < (3c+12)\epsilon$, and hence $\mathbb{E}\left[(\eta(X) - \hat{\eta}_n(X))^2\right] \to 0$ as $n \to \infty$.

  We first bound this expression by:

$$\begin{aligned}
\mathbb{E}\left[(\eta(X) - \hat{\eta}_n(X))^2\right] &\leq \mathbb{E}\left[\left(\eta(X) - \sum_{i=1}^{n} \eta(X_i) W_{ni}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} W_{ni}(X)\eta(X) - \sum_{i=1}^{n} \eta(X_i) W_{ni}(X)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} W_{ni}(X)\left(\eta(X) - \eta(X_i)\right)\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta(X_i))^2\right] \text{ by Jensen's inequality}
\end{aligned}$$

  We observe that any bounded measurable function is square integrable and so is in $L^2(\mu)$, and that continuous functions with bounded support are dense in $L^2(\mu)$ and are uniformly continuous.[17] Since η is bounded between zero and one and is measurable, this means there exists a uniformly continuous function η^* such that $\eta^*(x) \in [0,1]$ for all $x \in \mathbb{R}^d$ and

$$\mathbb{E}\left[(\eta(X) - \eta^*(X))^2\right] < \epsilon. \tag{2.13}$$

  We then find that (by applying Lemma 2.2.1):

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta(X_i))^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta^*(X) + \eta^*(X) - \eta^*(X_i) + \eta^*(X_i) - \eta(X_i))^2\right] \\
&\leq 3\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta^*(X))^2\right] + \\
&\quad 3\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2\right] +
\end{aligned}$$

$$3\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X_i) - \eta(X_i))^2\right]$$

For the first term, we see that

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta^*(X))^2\right] &= \mathbb{E}\left[(\eta(X) - \eta^*(X))^2 \sum_{i=1}^{n} W_{ni}(X)\right] \\ &= \mathbb{E}\left[(\eta(X) - \eta^*(X))^2\right] \\ &< \epsilon.\end{aligned}$$

For the second term $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2\right]$, we notice that since η^* is uniformly continuous, there exists an $a > 0$ such that if $\|X - X_i\| \leq a$, then $|\eta^*(X) - \eta^*(X_i)| < \sqrt{\epsilon}$. First apply this fact (by splitting the expectation into two disjoint sets, the part with $\|X - X_i\| \leq a$ and the part with $\|X - X_i\| > a$) and the linearity of integration. We then apply the fact that $(\eta^*(X) - \eta^*(X_i))^2 \leq 1$ always, and then use the second condition of the theorem to create a bound. We find that

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2(\mathbb{1}_{\{\|X_i - X\| \leq a\}} + \mathbb{1}_{\{\|X_i - X\| > a\}})\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2 \mathbb{1}_{\{\|X_i - X\| \leq a\}}\right] \\ &\quad + \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X) - \eta^*(X_i))^2 \mathbb{1}_{\{\|X_i - X\| > a\}}\right] \\ &\leq \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\sqrt{\epsilon}^2 \mathbb{1}_{\{\|X_i - X\| \leq a\}}\right] + \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] \\ &< \epsilon\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| \leq a\}}\right] + \epsilon \\ &\leq 2\epsilon.\end{aligned}$$

For the third term, we see that $(\eta^*(X_i) - \eta(X_i))^2$ is bounded above by 1 and is a measurable function of X_i (since both η and η^* are bounded), and so by the first assumption, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X_i) - \eta(X_i))^2\right] < c\epsilon + \epsilon_n.$$

We see that $\epsilon_n \to 0$ as $n \to \infty$, so we require n to be sufficiently large such that $\epsilon_n < \epsilon$. We then find that

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta^*(X_i) - \eta(X_i))^2\right] < (c+1)\epsilon.$$

Combining the results for these three terms, we find that for all sufficiently large n,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta(X_i))^2\right] &< 3\epsilon + 3(2\epsilon) + 3(c+1)\epsilon \\ &= (3c+12)\epsilon.\end{aligned}$$

Since $(3c+12)\epsilon$ can be made arbitrarily small by taking ϵ to be sufficiently small, it follows that $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)(\eta(X) - \eta(X_i)^2\right] \to 0$ as $n \to \infty$.

- We now show that $\mathbb{E}\left[(\hat{\eta}_n(X) - \eta_n(X))^2\right] \to 0$ as $n \to \infty$. We directly substitute in the definition of η_n and $\hat{\eta}_n$ into the expression and simplify. We obtain:

$$\begin{aligned}\mathbb{E}\left[(\hat{\eta}_n(X) - \eta_n(X))^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} \eta(X_i)W_{ni}(X) - \sum_{i=1}^{n} Y_i W_{ni}(X)\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n} W_{ni}(X)(\eta(X_i) - Y_i)\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ni}(X)W_{nj}(X)(\eta(X_i) - Y_i)(\eta(X_j) - Y_j)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}\left[W_{ni}(X)W_{nj}(X)(\eta(X_i) - Y_i)(\eta(X_j) - Y_j)\right]\end{aligned}$$

If $i \neq j$, we first apply the law of total expectation (in which we condition on $X, X_1, X_2, \ldots, X_n$ in the inner expectation), after which we notice that $W_{ni}(X)$, $W_{nj}(X)$, $(\eta(X_i) - Y_i)$, and $(\eta(X_j) - Y_j)$ are all conditionally independent with respect to $X, X_1, X_2, \ldots, X_n$, [1] which means that we can split the inner expectation, so we find:

$$\begin{aligned}
&\mathbb{E}\left[W_{ni}(X)W_{nj}(X)(\eta(X_i) - Y_i)(\eta(X_j) - Y_j)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[W_{ni}(X)W_{nj}(X)(\eta(X_i) - Y_i)(\eta(X_j) - Y_j)|X, X_1, X_2, \ldots, X_n\right]\right] \\
&= \mathbb{E}\left[\begin{smallmatrix}\mathbb{E}[W_{ni}(X)|X,X_1,X_2,\ldots,X_n]\times\mathbb{E}[W_{nj}(X)|X,X_1,X_2,\ldots,X_n]\times \\ \mathbb{E}[(\eta(X_i)-Y_i)|X,X_1,X_2,\ldots,X_n]\times\mathbb{E}[(\eta(X_j)-Y_j)|X,X_1,X_2,\ldots,X_n]\end{smallmatrix}\right]
\end{aligned}$$

We then notice that (since Y_i takes on values zero and one only, so we can replace the expected value of Y_i with the probability that $Y_i = 1$):

$$\begin{aligned}
&\mathbb{E}\left[\eta(X_i) - Y_i|X, X_1, X_2, \ldots, X_n\right] \\
&= \mathbb{E}\left[\eta(X_i)|X, X_1, X_2, \ldots, X_n\right] - \mathbb{E}\left[Y_i|X, X_1, X_2, \ldots, X_n\right] \\
&= \mathbb{E}\left[\mathbb{P}(Y_i = 1|X_i)|X, X_1, X_2, \ldots, X_n\right] - \mathbb{P}\left(Y_i = 1|X, X_1, X_2, \ldots, X_n\right) \\
&= \mathbb{P}\left(Y_i = 1|X, X_1, X_2, \ldots, X_n\right) - \mathbb{P}\left(Y_i = 1|X, X_1, X_2, \ldots, X_n\right) \\
&= 0
\end{aligned}$$

This implies that the expected value $\mathbb{E}\left[W_{ni}(X)W_{nj}(X)(\eta(X_i) - Y_i)(\eta(X_j) - Y_j)\right]$ (with $i \neq j$) is zero, since one of the factors in the expectation is zero (namely $\mathbb{E}\left[\eta(X_i) - Y_i|X, X_1, X_2, \ldots, X_n\right] = 0$) and all of the factors are finite. This means that the cross terms are all zero. Hence we have that the expectation is equal to the terms with $i = j$,

$$\mathbb{E}\left[(\hat{\eta}_n(X) - \eta_n(X))^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[W_{ni}(X)^2(\eta(X_i) - Y_i)^2\right]$$

[1] This holds since $W_{ni}(X)$ and $W_{nj}(X)$ are assumed to be functions of $X, X_1, X_2, \ldots, X_n$ only. If this condition is violated, the theorem fails, see counterexample 5.1.1.

$$
\begin{aligned}
&\leq \sum_{i=1}^{n} \mathbb{E}\left[W_{ni}(X)^2\right] \text{ since } (\eta(X_i) - Y_i)^2 \leq 1 \\
&\leq \sum_{i=1}^{n} \mathbb{E}\left[W_{ni}(X) \max_{1\leq i\leq n} W_{ni}(X)\right] \\
&= \mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(X) \sum_{i=1}^{n} W_{ni}(X)\right] \\
&= \mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(X)\right] \text{ since } \sum_{i=1}^{n} W_{ni}(X) = 1 \text{ always} \\
&\to 0 \text{ as } n \to \infty \text{ by the third condition.}
\end{aligned}
$$

∎

It can be shown that the k-NN learning rule on $\mathbb{R}^d$ (with any norm on $\mathbb{R}^d$) satisfies these conditions and so is universally consistent. This is what we will do in the next section.

2.3 Universal Consistency of k-NN

In this section, we prove that k-NN on the normed space $(\mathbb{R}^d, \|\cdot\|)$ is universally consistent. This is a known result, we explain the proof in detail as we will consider various extensions of this result later on. For the Euclidean norm, the result was first proven by Stone in [2]. A nice version of the proof for the Euclidean norm was presented in the book [14], the result for arbitrary norms appears to have been known to the authors of the book but was not proven. The full proof for arbitrary norms is done in [18].

We observe that the weight function for k-NN is:

$$
W_{ni}(X) = \begin{cases} \frac{1}{k} & \text{if } X_i \text{ is a } k\text{-nearest neighbour of } X \\ 0 & \text{otherwise} \end{cases} \tag{2.14}
$$

We notice that the weights are all nonnegative and sum to one. We then classify points using this weight function with equations (2.2) and (2.3).

An *inframetric space with a C-inframetric inequality* (Ω, ρ) is a nonempty set Ω together with a function $\rho : \Omega \times \Omega \to \mathbb{R}^+$ that satisfies:

1. $\rho(x, y) \geq 0$ and $\rho(x, y) = 0$ if and only if $x = y$ for all $x, y \in \Omega$.
2. $\rho(x, y) = \rho(y, x)$ for all $x, y \in \Omega$.
3. $\rho(x, z) \leq C \cdot \max\{\rho(x, y), \rho(y, z)\}$.

We easily see that any C-inframetric space satisfies a $2C$-weakened triangle inequality, that for all $x, y, z \in \Omega$, $\rho(x, z) \leq 2C(\rho(x, y) + \rho(y, z))$. As for metric spaces, we can define the notion of an open ball, open set, dense subset, separability, Borel σ-algebra, etc. for inframetric spaces. This is done for a more general family of symmetric kernels in [35]. First the open ball $B_r(y, \rho)$ is defined as the set $\{x \in \Omega \mid \rho(x, y) < r\}$ with the closed ball and sphere defined similarly ([35], part 1.1). Open sets, the notion of separability, and the Borel σ-algebra are then defined. The theory of measures is developed on such spaces.

Definition 2.3.1. The *support* of a measure μ is the set of points such that any open ball around any such point has nonzero measure, that is,

$$\text{Support}(\mu) = \{\mu \ : \ \forall r > 0, \ \mu(B_r(x)) > 0\}. \tag{2.15}$$

It can be easily shown that the support of a measure is always closed. We now prove a standard result about the support of a measure on an inframetric spaces (a sketch of the proof for $(\mathbb{R}^d, \|\cdot\|)$ can be found in [14] (Appendix 1, Lemma A.1), which works for any metric space). The result for spaces with a symmetric kernel satisfying a C-relaxed triangle inequality is proven in [35], Proposition 2.6.2.

Lemma 2.3.2. *The complement of the support has μ-measure zero in any separable C-inframetric space.*

Proof: We let A be the support of μ and T be a countable dense subset of Ω. By definition,

$$A^{\mathcal{C}} = \{x \in \Omega \ : \ \exists r > 0, \ \mu(B_r(x)) = 0\}. \tag{2.16}$$

We let $x \in A^{\mathcal{C}}$, and $r > 0$ be a radius around x such that $\mu(B_r(x)) = 0$. Without loss of generality we assume that r is rational. We see that there exists $y \in T$ such that $\rho(x, y) < \frac{r}{4C}$, and for any $z \in B_{r/(4C)}(y)$,

$$\begin{aligned} \rho(x, z) &\leq 2C(\rho(x, y) + \rho(y, z)) \\ &< 2C\left(\frac{r}{4C} + \frac{r}{4C}\right) \\ &= r. \end{aligned}$$

This means that $z \in B_r(X)$, and hence $\mu\left(B_{r/(2C)}(y)\right) = 0$.

By the above argument, we see that for all $x \in A^{\mathcal{C}}$, there exists $y_x \in T$ and rational $r_x > 0$ such that $x \in B_{r_x}(y_x)$ and $\mu(B_{r_x}(y_x)) = 0$. We define a family of such open balls

$$\mathcal{B} = \{B_r(y) \ : \ r > 0, \ r \in \mathbb{Q}, \ y \in T, \ \mu(B_r(y)) = 0\}. \tag{2.17}$$

It is clear that $\mathcal{B}$ is countable (since the countable union of countable sets is countable) and that every $x \in \mathcal{A}^{\mathcal{C}}$ is in $\mathcal{B}$, since it is in at least one of the open balls, and hence $A^{\mathcal{C}} \subseteq \cup_{B \in \mathcal{B}} B$. We then find (using subadditivity)

$$\begin{aligned} \mu(A^{\mathcal{C}}) &\leq \mu\left(\bigcup_{B \in \mathcal{B}} B\right) \\ &\leq \sum_{B \in \mathcal{B}} \mu(B) \\ &= \sum_{B \in \mathcal{B}} 0 \\ &= 0. \end{aligned}$$

∎

Lemma 2.3.3. *Let $(A_n)_{n=1}^{\infty}$ be a sequence of random variables that converges almost surely to zero as n approaches infinity. We then have that the supremum of the tail also converges almost surely to zero, $\sup_{m\geq n} A_m \to 0$ with probability one as $n \to \infty$.*

Proof: We first prove this for deterministic sequences. Let $(a_n)_{n=1}^{\infty}$ be a deterministic sequence that converges to zero. For any $\epsilon > 0$, there exists $N \geq 1$ such that for all $n \geq N$, $|a_n| < \epsilon/2$. This means that $\sup_{m\geq N} a_m \leq \epsilon/2 < \epsilon$. We observe the sequence $\left(\sup_{m\geq N} a_m\right)_{n=1}^{\infty}$ is monotone decreasing. Hence we have that for all $n \geq N$, $\sup_{m\geq n} a_m \leq \sup_{m\geq N} a_m < \epsilon$. It follows that for all $\epsilon > 0$, there exists $N \geq 1$ such that for all $n \geq N$, $\sup_{m\geq n} a_m < \epsilon$, and hence $\sup_{m\geq n} a_m \to 0$ as $n \to \infty$.

If A_n converges almost surely to zero, we have that $A_n(\omega) \to 0$ as $n \to \infty$ for all $\omega \in \Omega_0$, for some subset Ω_0 of the probability space with $\mathbb{P}(\Omega_0) = 1$. Since the result holds for deterministic sequences, we have that for all points $\omega \in \Omega_0$, $\sup_{m\geq n} A_m(\omega) \to 0$ as $n \to \infty$, and hence $\sup_{m\geq n} A_m$ converges to zero almost surely as $n \to \infty$. ∎

This following result is a generalization of a result originally proved by Cover and Hart in [20], see also [14] (Lemma 5.1). Our new result extends the result to inframetric spaces.

Lemma 2.3.4. *Suppose we have a separable inframetric space (Υ, ρ) with probability measure $\mathbb{P}$. Given iid points $X, X_1, X_2, \ldots, X_n$, let $X_{(1,\rho)}, X_{(2,\rho)}, \ldots, X_{(n,\rho)}$ be the points in increasing distance from X with respect to the metric ρ and let $a > 0$ be a constant. Then as $n \to \infty$, for any sequence $(k_n)_{n=1}^{\infty}$ such that $\frac{k_n}{n} \to 0$,*

$$\mathbb{P}(\rho(X_{(k_n)}, X) > a) \to 0. \tag{2.18}$$

Proof: We first notice that for all $x \in \text{Support}(\mathbb{P})$ and $\epsilon > 0$,

$$\rho(X_{(k_n)}(x), x) \geq \epsilon \Leftrightarrow \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B_\epsilon(x)\}} < k_n$$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B_\epsilon(x)\}} < \frac{k_n}{n}.$$

We then notice that $k_n/n \to 0$ as $n \to \infty$ by assumption and $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B_\epsilon(x)\}} \to \mathbb{P}(B_\epsilon(x))$ almost surely as $n \to \infty$ by the strong law of large numbers, and by assumption $\mu(B_\epsilon(x)) > 0$ since x is in the support of μ. It follows that $\rho(X_{(x)} - x) \to 0$ as $n \to \infty$ with probability one.

We define $X_{(k,n)}(x)$ to be the k^{th} order statistic of x from the sample $X_1, X_2, \ldots, X_n$ of size n in the ρ distance (this notation makes the sample size clear when we discuss the order statistics). By the above argument we have that for any x in the support of $\mathbb{P}$, we have $\rho(X_{(k_n,n)}(x) - x) \to 0$ as $n \to \infty$ with probability one. From Lemma 2.3.3 we have that $\sup_{m \geq n} \rho(X_{(k_m,m)}(x) - x) \to 0$ as $n \to \infty$ with probability one as well.

We then notice that for the random variable X, by Lemma 2.3.2,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{m \geq n} \rho(X_{(k_m,m)}(X), X) > \epsilon\right) \\
&= \mathbb{P}(X \in \text{Support}(\mathbb{P}))\mathbb{P}\left(\sup_{m \geq n} \rho(X_{(k_m,m)}(X), X) > \epsilon \middle| X \in \text{Support}(\mathbb{P})\right) + \\
&\quad \mathbb{P}(X \notin \text{Support}(\mathbb{P}))\mathbb{P}\left(\sup_{m \geq n} \rho(X_{(k_m,m)}(X), X) > \epsilon \middle| X \notin \text{Support}(\mathbb{P})\right) \\
&= \mathbb{P}\left(\sup_{m \geq n} \rho(X_{(k_m,m)}(X), X) > \epsilon \middle| X \in \text{Support}(\mathbb{P})\right).
\end{aligned}$$

Since the sequence $\sup_{m \geq n} \rho(X_{(k_m,m)}(x), x)$ is nonnegative, monotone decreasing, and converges to zero almost everywhere if $X \in \text{Support}(\mathbb{P})$, by the Monotone Convergence Theorem (applied to the expectations of the indicator functions of the events $\sup_{m \geq n} \rho(X_{(k_m,m)}(X), X) > \epsilon$) we find that the conditional probability $\mathbb{P}\left(\sup_{m \geq n} \rho(X_{(k_m,m)}(X), X) > \epsilon \middle| X \in \text{Support}(\mathbb{P})\right) \to 0$ as $n \to \infty$. Since $0 \leq \rho(X_{(k_n,n)}(X), X) \leq \sup_{m \geq n} \rho(X_{(k_m,m)}(X), X)$, we have that as $n \to \infty$,

$$\mathbb{P}\left(\rho(X_{(k_n,n)}(X), X) > \epsilon\right) \leq \mathbb{P}\left(\sup_{m \geq n} \rho(X_{(k_m,m)}(X), X) > \epsilon \middle| X \in \text{Support}(\mathbb{P})\right) \to 0.$$

∎

We recall that the *weights* $W_{ni}(X)$ (with $1 \leq i \leq n$) are functions of $X, X_1, X_2, \ldots, X_n$ that are nonnegative and sum to one, and from equation (2.2) if the sum of the weights $W_{ni}(X)$ multiplied by the corresponding Y_i is at least $1/2$, we assign label one, otherwise we assign label zero. For k-NN, we recall that the weights are $1/k$ for the k-nearest points to the query, and are zero otherwise. The following result follows easily from Lemma 2.3.4 and is proven in [14] (inside the proof of Theorem 6.4).

Lemma 2.3.5. *If we let $W_{ni}(X)$ be the weights in the k-NN learning rule for the normed space $(\mathbb{R}^d, \|\cdot\|)$, then*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] \to 0 \text{ as } n \to \infty. \tag{2.19}$$

Proof: We see that $\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}$ is bounded above by one, is nonnegative, and is nonzero if and only if the k^{th} nearest point to X has a distance of at most a. By Lemma 2.3.4, the probability of this goes to zero as n goes to infinity, and hence the expected value $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] \to 0$ as $n \to \infty$. ∎

Lemma 2.3.6. *For any norm $\|\cdot\|$ on $\mathbb{R}^d$ and radius $\delta > 0$, the unit sphere $S_1(\mathbf{0}, \|\cdot\|)$ can be covered by c balls of radius δ each in the $\|\cdot\|$ norm, that is,*

$$S_1(\mathbf{0}, \|\cdot\|) \subseteq \bigcup_{i=1}^{c} B_\delta(\boldsymbol{x}_i, \|\cdot\|). \tag{2.20}$$

Proof: This holds since $S_1(\mathbf{0}, \|\cdot\|)$ is a bounded subset of $\mathbb{R}^d$. ∎

The result that k-NN satisfies the first condition in Stone's theorem is called *Stone's Lemma*. We now present the proof of Stone's lemma for any norm on $\mathbb{R}^d$

(the *Generalized Stone's Lemma*, as originally Stone's Lemma was only proved for the Euclidean norm in [2], and a modified version of the proof was presented in [14], with cones of angle $\pi/6$). The result for arbitrary norms on $\mathbb{R}^d$ is given as an exercise in [14] (Chapter 5, Problem 5.1). The first published proof of the general result (for any norm on $\mathbb{R}^d$) that I am aware of is in [18] (Lemma 2.2.9).

Lemma 2.3.7. *Let c be the number of subsets such that the unit sphere $S_1(\mathbf{0}, \|\cdot\|)$ can be covered by c balls of radius $1/4$ each. Then there exist c subsets $S_1, S_2, \ldots, S_c$ (with each of them containing the zero vector) covering $\mathbb{R}^d$ such that in every subset S_q (with $1 \leq q \leq c$), if $\boldsymbol{x}, \boldsymbol{y} \in S_q$ with $\|\boldsymbol{x}\| \leq \|\boldsymbol{y}\|$ and $\boldsymbol{x} \neq \mathbf{0}$, then $\|\boldsymbol{y} - \boldsymbol{x}\| < \|\boldsymbol{y}\|$.*

Proof: We first see by Lemma 2.3.6 that there exists a finite covering of c of the unit sphere $S_1(\mathbf{0}, \|\cdot\|)$ by open balls of radius $1/4$. We let the points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_c$ be the centres of the balls of such a covering. For each open ball $B_{1/4}(\boldsymbol{x}_i, \|\cdot\|)$ (with $1 \leq i \leq c$), we define the set A_i by having $\mathbf{0} \in A_i$ always and for all $\boldsymbol{x} \neq \mathbf{0}$,

$$\boldsymbol{x} \in A_i \Leftrightarrow \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \in B_{1/4}(\mathbf{x}_i, \|\cdot\|). \tag{2.21}$$

We then see that the sets $A_1, A_2, \ldots, A_c$ cover $\mathbb{R}^d$, since the zero vector is in all of the sets and for every nonzero vector $\boldsymbol{x}$, $\|\boldsymbol{x}/\|\boldsymbol{x}\|\|$ lies on the unit sphere which is covered by the above set of open balls and so $\boldsymbol{x}$ is in at least one A_i.

We then see that if $\boldsymbol{x}, \boldsymbol{y} \in A_i$, then

$$\begin{aligned}
\left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \right\| &= \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \boldsymbol{x}_i + \boldsymbol{x}_i - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \right\| \\
&\leq \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \boldsymbol{x}_i \right\| + \left\| \boldsymbol{x}_i - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \right\| \\
&< \frac{1}{4} + \frac{1}{4} \\
&= \frac{1}{2}.
\end{aligned}$$

It then follows that

$$\left\| \frac{\boldsymbol{y}\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} - \boldsymbol{x} \right\| = \left\| \|\boldsymbol{x}\| \left(\frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} - \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \right) \right\|$$

$$\begin{aligned} &= \|\boldsymbol{x}\| \left\| \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} - \frac{\boldsymbol{y}}{\|\boldsymbol{y}\|} \right\| \\ &< \frac{\|\boldsymbol{x}\|}{2}. \end{aligned}$$

From this, we are able to find that

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{x}\| &= \left\| \boldsymbol{y} - \frac{\boldsymbol{y}\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} + \frac{\boldsymbol{y}\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} - \boldsymbol{x} \right\| \\ &\leq \left\| \boldsymbol{y} - \frac{\boldsymbol{y}\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \right\| + \left\| \frac{\boldsymbol{y}\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} - \boldsymbol{x} \right\| \\ &< \left\| \boldsymbol{y} - \frac{\boldsymbol{y}\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \right\| + \frac{\|\boldsymbol{x}\|}{2} \\ &= \left\| \left(1 - \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \right) \boldsymbol{y} \right\| + \frac{\|\boldsymbol{x}\|}{2} \\ &= \left(1 - \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{y}\|} \right) \|\boldsymbol{y}\| + \frac{\|\boldsymbol{x}\|}{2} \\ &= \|\boldsymbol{y}\| - \frac{\|\boldsymbol{x}\|}{2} \\ &< \|\boldsymbol{y}\|. \end{aligned}$$

∎

The following result has been proven in [18] (Theorem 2.2.8).

Lemma 2.3.8. *Suppose we have the finite dimensional normed vector space $(\mathbb{R}^d, \|\cdot\|)$, we let $f : \mathbb{R}^d \to \mathbb{R}$ be any nonnegative measurable function with finite expected value (in terms of the measure μ), we let $1 \leq k \leq n$, and we define c to be a constant such that $\mathbb{R}^d$ can be partitioned into a finite number of subsets $S_1, S_2, \ldots, S_c$, such that if $1 \leq q \leq c$, if $\boldsymbol{x}, \boldsymbol{y} \in S_q$, $\boldsymbol{x} \neq \boldsymbol{0}$, and $\|\boldsymbol{x}\| \leq \|\boldsymbol{y}\|$, then $\|\boldsymbol{y} - \boldsymbol{x}\| < \|\boldsymbol{y}\|$. We let $X, X_1, X_2, \ldots, X_n$ be iid random variables on $\mathbb{R}^d \times \{0, 1\}$ with probability distribution μ. If we let $W_{ni}(X)$ be the weights in k-NN with the $\|\cdot\|$ norm, we find that*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i) \right] \leq c\mathbb{E}\left[f(X) \right]. \tag{2.22}$$

Proof: Given a query X, we define the subsets $S'_1, S'_2, \ldots, S'_c$ as

$$X_i \in S'_q \Leftrightarrow X - X_i \in S_q. \tag{2.23}$$

We see that since $S_1, S_2, \ldots, S_c$ cover $\mathbb{R}^d$ and $X_i - X$ is a vector in $\mathbb{R}^d$, the new subsets cover $\mathbb{R}^d$, that is, $\mathbb{R}^d$ is covered by $S'_1, S'_2, \ldots, S'_c$. We let $S'_q \in \{S'_1, S'_2, \ldots, S'_c\}$ be one the subsets. In the subset S'_q, we mark the k points closest to X in the $\|\cdot\|$ norm among $\{X_1, X_2, \ldots, X_n\} \cap S'_q$ (if there are fewer than k such points, we take all of them, and we break distance ties by generating independent uniform random variables $U_1, U_2, \ldots, U_n$ and taking the point such that U_i is larger, as discussed previously). We see that the number of points that are marked in the subset S'_q is at most k. If a point $X_i \in S'_q$ is not marked, then there must exist at least k points in S'_q that are either closer to X than X_i in the $\|\cdot\|$ norm or have the same distance but the independent random variable U_i is larger. Let $X_j \in S'_q$ be such a point. We need to show that X_j is closer to X_i than X is (in the case of ties, the tiebreaking variables $U_1, U_2, \ldots, U_n$ define which point is "closer"). Since $X_i, X_j \in S'_q$, $X - X_i, X - X_j \in S_q$. There are now two possible cases:

(i) If $X_j \neq X$, $X - X_j \neq \mathbf{0}$, and by assumption $\|X - X_j\| \leq \|X - X_i\|$. This implies (by the definition of the subset S_j) that

$$\begin{aligned} \|X_i - X_j\| &= \|(X_i - X) + (X - X_j)\| \\ &= \|(X_i - X) - (X_j - X)\| \\ &< \|X_i - X\| \end{aligned}$$

and so X_j is closer to X_i than X is.

(ii) Otherwise, if $X_j = X$, then $U_j > U_i$ (these being the tie-breaking variables discussed earlier), so in our tie-breaking rule (for the nearest neighbour) we select X_j before X from the list $\{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_j \ldots, X_n\} \cap S'_q$ (where X takes the place of X_i in the list of points, so the same U_i variable is used for X here).

This holds for all the other k points in S'_q that are nearest X, and so X is not a k nearest neighbour of X_i.

Hence we see that if X_i is not marked, X is not a k-nearest neighbour of X_i. Equivalently, the set of k-nearest neighbours of X is a subset of the set of points that are marked. The number of points that are marked is at most ck, since there are c subsets and each subset contains at most k marked points. We see that

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[W_{ni}(X) f(X_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{k} \mathbb{1}_{\{X_i \text{ is a } k\text{-nearest neighbour of } X \text{ in the } \|\cdot\| \text{ norm among } X_1, X_2, \ldots, X_n\}} f(X_i)\right] \\
&= \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{X \text{ is a } k\text{-nearest neighbour of } X_i \text{ in the } \|\cdot\| \text{ norm among } X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}} f(X)\right] \\
&= \frac{1}{k} \mathbb{E}\left[f(X) \sum_{i=1}^{n} \mathbb{1}_{\{X \text{ is a } k\text{-nearest neighbour of } X_i \text{ in the } \|\cdot\| \text{ norm among } X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}}\right] \\
&\leq \frac{1}{k} \mathbb{E}\left[f(X) \sum_{i=1}^{n} \mathbb{1}_{\{X_i \text{ is marked}\}}\right] \\
&\leq \frac{1}{k} \mathbb{E}\left[f(X)(k)c\right] \\
&= c\mathbb{E}\left[f(X)\right].
\end{aligned}
$$

∎

The result that k-NN is universally consistent on the normed space $(\mathbb{R}^d, \|\cdot\|)$ appears to have been known to the authors of [14], where Stone's lemma (which is the first condition for Stone's theorem) for arbitrary norms on $\mathbb{R}^d$ is given as an exercise earlier (Chapter 5, Problem 5.1), universal consistency is proven for the Euclidean norm, and the proof of universal consistency for a fixed ℓ^p norm (with $1 \leq p \leq \infty$) is left as an exercise (Chapter 11, Problem 11.4, which recommends proving universal consistency by checking the conditions of Stone's theorem; the

statement of the problem in the book includes the ℓ^p quasinorms with $0 < p < 1$ (by giving the problem for $0 < p \leq \infty$), which I think is a misprint as the geometric Stone's lemma does not hold for the ℓ^p quasinorms with $0 < p < 1$ as we show in Example 3.6.1 and no alternative approach for quasinorms is stated in the book). However, no proof for arbitrary norms is provided in the book. The first published proof for arbitrary norms that I am aware of is in [18] (Theorem 2.2.1).

Theorem 2.3.9. *The k-NN classifier on the normed space $(\mathbb{R}^d, \|\cdot\|)$ with $k \to \infty$ as $n \to \infty$ and $\frac{k}{n} \to 0$ as $n \to \infty$ is universally consistent.*

Proof: We show that k-NN satisfies the conditions of Stone's theorem:

1. The first condition holds because the unit sphere $S_1(\mathbf{0}, \|\cdot\|)$ can be covered by finitely many balls of radius $1/4$ each (by Lemma 2.3.7) and hence Lemma 2.3.8 applies.

2. The second condition holds by Lemma 2.3.5.

3. The third condition holds since $k \to \infty$ as $n \to \infty$, so $\frac{1}{k} \to 0$ as $n \to \infty$.

∎

Invariance of k-NN under Strictly Increasing Transformations

We now make the observation that we can apply any strictly increasing transformation to our distance function in k-NN and k-NN will generate the same predictions for each point.

Lemma 2.3.10. *Let X be the query and $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be the sample. We define Y to be the label predicted for X by k-NN with the function d as the distance. If we let h be a strictly increasing function and Y' to be the label predicted for X by*

k-NN with the function $h \circ d$ (the composition of the functions d followed by h) as the distance, then $Y' = Y$.

Proof: Suppose a point X_i is a k-nearest neighbour of X with the distance function d, so there are fewer than k sample points closer to X than X_i under d. If X_j is a point such that $d(X, X_j) > d(X, X_i)$, then $h \circ d(X, X_j) > h \circ d(X, X_i)$, so X_j remains further away than X_i. If $d(X, X_j) = d(X, X_i)$, then $h \circ d(X, X_j) = h \circ d(X, X_i)$, so the distance tie remains, which is broken by comparing the tiebreaking variables, which remain the same, so X_i remains as a k-nearest neighbour. This means if a point X_i has fewer than k points closer than it to X under d, this remains true under $h \circ d$, hence the k-nearest neighbours remain the same under $h \circ d$. ∎

This result allows us to apply a strictly increasing function to the distance kernel used for k-NN (that is, to find the distance between points $\boldsymbol{x}, \boldsymbol{y}$ for k-NN, we compute $f \circ d(x, y)$) and keep the same results with k-NN. This can be useful in reducing the computation time required for classification. For instance, when using the Euclidean distance for k-NN, we instead compute the Euclidean distance squared, which is $\sum_{i=1}^{d} x_i^2$, instead of the square root of this, and we obtain the same results (this eliminates the need for us to calculate the square root, which reduces our computation time required for k-NN). This also eliminates the need for us to include scaling coefficients or shift factors in our distance function in some cases.

Chapter 3

k-NN with a Sequence of Random Norms

Suppose in the k-NN learning rule, we have a sequence of random norms from some family of norms, instead of a single norm. We show that under certain conditions, the resultant learning rule is universally consistent. A result of a form similar to ours (without the independence assumptions we make for the sequence of random norms) is found in [14], unfortunately, as we explain below the proof in the book is incomplete. We do not know if the result is correct after we remove the independence assumption.

3.1 Families of Norms

We now define a partial ordering $\preceq$ on the family of all norms $\mathcal{F}$ on a vector space V. For two norms $\|\cdot\|_A, \|\cdot\|_B \in \mathcal{F}$,

$$\|\cdot\|_B \preceq \|\cdot\|_A \text{ if and only if } \forall \boldsymbol{v} \in V \ \|\boldsymbol{v}\|_B \leq \|\boldsymbol{v}\|_A. \tag{3.1}$$

Lemma 3.1.1. *The relation $\preceq$ in* 3.1 *is a partial ordering on the family of norms $\mathcal{F}$ on a vector space V.*

Proof: This is easily seen by verifying the conditions of a partial order. ∎

Lemma 3.1.2. *If $p, q \in (0, \infty) \cup \{\infty\}$ with $p \geq q$, then for all $\boldsymbol{v} \in \mathbb{R}^d$,*

$$\|\cdot\|_q \succeq \|\cdot\|_p. \tag{3.2}$$

Lemma 3.1.3. *If $\|\cdot\|_A$ and $\|\cdot\|_B$ are norms on $\mathbb{R}^d$ where $\|\cdot\|_B \preceq \|\cdot\|_A$, then for any point $\mathbf{x} \in \mathbb{R}^d$ and radius $r > 0$, the open balls and closed balls in norm $\|\cdot\|_A$ are smaller than those in norm $\|\cdot\|_B$:*

$$B_r(\mathbf{x}, \|\cdot\|_A) \subseteq B_r(\mathbf{x}, \|\cdot\|_B) \tag{3.3}$$

$$B_r^-(\mathbf{x}, \|\cdot\|_A) \subseteq B_r^-(\mathbf{x}, \|\cdot\|_B) \tag{3.4}$$

Lemma 3.1.4. *Let $\|\cdot\|_A, \|\cdot\|_B$ be two norms on $\mathbb{R}^n$ and $\epsilon > 0$. If a set $V \subseteq \mathbb{R}^d$ can be covered by finitely many open ϵ-balls in the norm $\|\cdot\|_A$, then it can be covered by finitely many open ϵ-balls in the norm $\|\cdot\|_B$.*

Proof: By the equivalence of norms on $\mathbb{R}^d$, there exists a constant $C \geq 1$ such that

$$\frac{1}{C}\|\boldsymbol{v}\|_A \leq \|\boldsymbol{v}\|_B \leq C\|\boldsymbol{v}\|_A.$$

Any subset $V \subseteq \mathbb{R}^d$ that can be covered by finitely many ϵ-balls in the $\|\cdot\|_A$ norm is bounded in that norm, and since this is a subset of $\mathbb{R}^d$, is totally bounded (pre-compact). This means that there exists a finite set of points S such that the balls of radius ϵ/C in the $\|\cdot\|_A$ norm around these points cover V. If we let $\boldsymbol{x} \in V$ and $\boldsymbol{y} \in S$ such that $\|\boldsymbol{x} - \boldsymbol{y}\| < \epsilon/C$. We then find that

$$\|\boldsymbol{x} - \boldsymbol{y}\|_B \leq C\|\boldsymbol{x} - \boldsymbol{y}\|_A < C\frac{\epsilon}{C} = \epsilon.$$

∎

3.2 Consistency of k-NN with a Family of Norms

In this section, we define $\mathcal{N}$ to be a family of norms on $\mathbb{R}^d$, we define the conditions $\mathcal{N}$ must satisfy in each lemma and theorem. As usual, we assume that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ is an iid labelled sample and let (X, Y) be the query and its response, which is independent from the sample and follows the same distribution. We let V be an arbitrary random variable independent from both the query and the sample. It is possible to assume that the tiebreaking variables $U_1, U_2, \ldots, U_n$ are contained in V (even if we require V to be a classical real-values random variable, we can simply use a Borel isomorphism from $\mathbb{R}^n$ to $\mathbb{R}$ to combine the n random variables into a single random variable that is independent of the labelled sample and the query).

Lemma 3.2.1. *Let $\mathcal{N}$ be a family of norms on $\mathbb{R}^d$ such that there exist two norms $\|\cdot\|_U$ and $\|\cdot\|_L$, such that for all $\rho \in \mathcal{N}$, $\|\cdot\|_L \preceq \rho \preceq \|\cdot\|_U$. There exists a finite number c such that for any norm $\rho \in \mathcal{N}$ the unit sphere $S_1(\mathbf{0}, \rho)$ can be covered by c open balls of radius $1/4$.*

Proof: Let $S_1(\mathbf{0}, \rho)$ be the unit sphere in norm ρ. It is clear that the unit sphere is a subset of the closed unit ball, $S_1(\mathbf{0}, \rho) \subseteq B_1^-(\mathbf{0}, \rho)$, and by Lemma 3.1.3 it follows that $B_1^-(\mathbf{0}, \rho) \subseteq B_1^-(\mathbf{0}, \|\cdot\|_L)$, so that $S_1(\mathbf{0}, \rho) \subseteq B_1(\mathbf{0}, \|\cdot\|_L)$.

We have that $B_1^-(\mathbf{0}, \|\cdot\|_L)$ is compact in the $\|\cdot\|_L$ norm, so it is bounded, by Lemma 3.1.4 it is also bounded in the $\|\cdot\|_U$ norm, hence there is a finite subcover of c open balls of $1/4$ radius in the $\|\cdot\|_U$ norm, that is there exists a set of points $\mathbf{x_1}, \ldots, \mathbf{x_c}$ such that

$$B_1^-(\mathbf{0}, \|\cdot\|_L) \subseteq \bigcup_{i=1}^{c} B_{1/4}^-(\mathbf{x_i}, \|\cdot\|_U).$$

By Lemma 3.1.3, $B_{1/4}(\mathbf{x}_i, \|\cdot\|_U) \subseteq B_{1/4}(\mathbf{x}_i, \rho)$, which means that

$$B_1^-(\mathbf{0}, \|\cdot\|_L) \subseteq \bigcup_{i=1}^{c} B_{1/4}(\mathbf{x}_i, \|\cdot\|_U) \subseteq \bigcup_{i=1}^{c} B_{1/4}(\mathbf{x}_i, \rho). \tag{3.5}$$

Since $S_1(\mathbf{0}, \rho) \subseteq B_1^-(\mathbf{0}, \|\cdot\|_L)$, it follows that the set of open balls of radius $1/4$ around $\mathbf{x}_1, \ldots, \mathbf{x}_c$ (in any of the norms in $\mathcal{N}$) covers $S_1(\mathbf{0}, \rho)$. ∎

Lemma 3.2.2. *Suppose $\mathcal{N}$ is a family of norms that is bounded above by some norm $\|\cdot\|_U$ and below by another norm $\|\cdot\|_L$. We let W_{ni} be the weight function for k-NN with the norm $\rho_n \in \mathcal{N}$ for each n, with ρ_n being independent of the sample and the query. For every nonnegative measurable function $f : \mathbb{R}^d \to \mathbb{R}$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}\left[f(X)\right]. \tag{3.6}$$

Proof: By Lemma 3.2.1, we have that there exists a constant c such that there are points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_c$ so that for any norm $\rho \in \mathcal{N}$, the unit sphere $S_1(\mathbf{0}, \rho)$ is covered by $B_{1/4}(\boldsymbol{x}_1, \rho), B_{1/4}(\boldsymbol{x}_2, \rho), \ldots, B_{1/4}(\boldsymbol{x}_c, \rho)$. By Lemma 2.3.7, there exists a corresponding set of cones $S_1, S_2, \ldots, S_c$ that cover $\mathbb{R}^d$, such that within any cone S_q, if $\boldsymbol{x}, \boldsymbol{y} \in S_q$ with $0 < \rho(\boldsymbol{x}) \leq \rho(\boldsymbol{y})$, then $\rho(\boldsymbol{y} - \boldsymbol{x}) < \rho(\boldsymbol{y})$ for any norm $\rho \in \mathcal{N}$. A point is marked in a given norm if it is one of the k-nearest neighbours of X among points in the sample that are in a given cone, that is, for each cone S_q, we mark the k points among $X_1, \ldots, X_n$ in the cone S_i that are closest to X (if there are fewer than k points in a given cone, we mark all of them, and we break distance ties by comparing independent uniform random variables $U_1, U_2, \ldots, U_n$ as discussed previously). It follows by and the argument in Lemma 2.3.8 that for any fixed point and sample in $\mathbb{R}^d$ and norm $\rho \in \mathcal{N}$, the set of k-nearest neighbours of a point in a sample is a subset of the set of points that are marked and at most ck points are marked. We have that ρ_n is always a norm in $\mathcal{N}$ and is independent of the sample and the query. We can expand $W_{ni}(X)$ according to our definition and exchange X and X_i in the expectation (since they are iid and are independent of everything else)

to find that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{k} \mathbb{1}_{\{X_i \text{ is } \rho_n \text{ } k\text{-NN of } X \text{ among } X_1,\ldots,X_{i-1},X_i,X_{i+1},\ldots,X_n\}} f(X_i)\right] \\
&= \frac{1}{k} \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{\{X_i \text{ is } \rho_n \text{ } k\text{-NN of } X \text{ among } X_1,\ldots,X_{i-1},X_i,X_{i+1},\ldots,X_n\}} f(X_i)\right] \\
&= \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{X \text{ is } \rho_n \text{ } k\text{-NN of } X_i \text{ among } X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}} f(X)\right] \\
&\leq \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{X \text{ is marked in } \rho_n \text{ norm}\}} f(X)\right] \\
&\leq \frac{1}{k} ck \mathbb{E}\left[f(X)\right] \\
&= c\mathbb{E}\left[f(X)\right].
\end{aligned}$$

∎

Lemma 3.2.3. *Let $\mathcal{N}$ be a family of norms in $\mathbb{R}^d$ such that for some norm $\|\cdot\|$ and constant $C \geq 1$, $\forall \rho \in \mathcal{N}$, $\frac{1}{C}\|\cdot\| \preceq \rho \preceq C\|\cdot\|$. Given a random point X, let $X_{(1,\|\cdot\|)}, \ldots, X_{(n,\|\cdot\|)}$ be the points in increasing distance from X with respect to the norm $\|\cdot\|$. For any $\rho \in \mathcal{N}$, if W_{ni} are the weights in k-NN with ρ as the norm, then for any $a > 0$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > a\}}\right] \leq \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{a}{C^2}\right). \tag{3.7}$$

Proof: We first notice that the function $\sum_{i=1}^{n} W_{ni}(X)$ in the expectation is bounded above by one, hence the expectation is bounded above by the probability that the inner expression is nonzero,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > a\}}\right] \leq \mathbb{P}\left(\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > a\}} \neq 0\right). \tag{3.8}$$

Suppose that $\left\|X_{(k,\|\cdot\|)} - X\right\| \leq a/C^2$. Since $\rho(\boldsymbol{x}) \leq \|\boldsymbol{x}\|$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and for any $i \leq k$, $\left\|X_{(i,\|\cdot\|)} - X\right\| \leq \left\|X_{(k,\|\cdot\|)} - X\right\|$,

$$\begin{aligned}\rho\left(X_{(i,\|\cdot\|)} - X\right) &\leq C\left\|X_{(i,\|\cdot\|)} - X\right\| \\ &\leq C\left\|X_{(k,\|\cdot\|)} - X\right\| \\ &\leq C\frac{a}{C^2} \\ &= \frac{a}{C}.\end{aligned}$$

This means there exist at least k points $X_{(1,\|\cdot\|)}, X_{(2,\|\cdot\|)}, \ldots, X_{(k,\|\cdot\|)}$ such that $\rho\left(X_{(i,\|\cdot\|)} - X\right) \leq a/C$. For any point X_j such that $\|X_j - X\| > a$, we have that $a < \|X_j - X\| \leq C\rho(X_j - X)$, and hence $\rho(X_j - X) > a/C$. This means that X_j cannot be a k-nearest neighbour of X in the ρ distance, as there are at least k points in the sample whose ρ distance to X is less than or equal to a/C. It follows that, if $\left\|X_{(k,\|\cdot\|)} - X\right\| \leq a/C^2$, then the interior of the expectation $\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}$ (from equation (3.8)) is zero, as no point X_i can be a k-nearest neighbour of X in the ρ distance and satisfy $\|X_i - X\| > C^2 a$ simultaneously.

Hence we find that if the term inside the expectation in equation (3.8) is nonzero, then $\left\|X_{(k,\|\cdot\|)} - X\right\| > a/C^2$ must hold. We conclude

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] &\leq \mathbb{P}\left(\sum_{i=1}^n W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}} \neq 0\right) \\ &\leq \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{a}{C^2}\right).\end{aligned}$$

∎

Corollary 3.2.4. *Let $(\rho_n)_{n=1}^\infty$ be a sequence of random norms independent of the sample $X_1, X_2, \ldots, X_n$ and query X, and let W_{ni} be the weights in the k-NN classifier with ρ_n as the norm. For any $a > 0$, we have that*

$$\mathbb{E}\left[W_{ni}(X)\mathbb{1}_{\{\|X_i - X\} > a\|\}}\right] \to 0 \text{ as } n \to \infty. \tag{3.9}$$

Proof: We apply the law of total expectation (conditioning on the norm ρ), we then use the fact that the norm ρ is independent of $X, X_1, X_2, \ldots, X_n$ and apply Lemma 3.2.3 to find

$$\begin{aligned}
&\mathbb{E}\left[W_{ni}(X)\mathbb{1}_{\{\|X_i-X\}>a\|\}}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[W_{ni}(X)\mathbb{1}_{\{\|X_i-X\|>a\}}\middle|\rho\right]\right] \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\|X_i - X\| > \frac{a}{C^2}\middle|\rho\right)\right] \\
&= \mathbb{P}\left(\|X_i - X\| > \frac{a}{C^2}\right).
\end{aligned}$$

We then notice that the last term goes to zero by Lemma 2.3.4. ∎

We are now able to prove our result that k-NN with a sequence of random norms (chosen independently of the sample and query) from a family of norms $\mathcal{N}$ satisfying certain boundedness condition is universally consistent. An example of a family $\mathcal{N}$ that satisfies our conditions is the family of all ℓ^p norms (with $1 \leq p \leq \infty$). For each $n \geq 1$, ρ_n is a random norm from $\mathcal{N}$, with n being the sample size. This allows us to pick the random norm from $\mathcal{N}$ differently as the sample size changes (as long as we keep independence from the sample and query).

Theorem 3.2.5. *Let $\mathcal{N}$ be a family of norms on $\mathbb{R}^d$ such that there exist norms $\|\cdot\|_L, \|\cdot\|_U$ where $\forall \rho \in \mathcal{N}\ \|\cdot\|_L \preceq \rho \preceq \|\cdot\|_U$. For any sequence of random norms $(\rho_n)_{n=1}^{\infty}$ in $\mathcal{N}$ that are independent of the query and the sample, k-NN with this sequence of norms is universally consistent.*

Proof: We verify that the k-NN learning rule with the norm chosen from $\mathcal{N}$ by the function ρ_n at each step satisfies the conditions for Stone's Theorem:

1. By Lemma 3.2.2, we have that for every nonnegative measurable function $f : \mathbb{R}^d \to \mathbb{R}$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}\left[f(X)\right].$$

2. The second condition is satisfied as shown by Corollary 3.2.4.

3. This follows directly from the fact that in the k-NN classifier, as $n \to \infty$, $k_n \to \infty$, so that $1/k_n \to 0$, and hence

$$\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(X)\right] = \mathbb{E}\left[1/k_n\right] = 1/k_n \to 0 \text{ as } n \to \infty.$$

It follows from Stone's theorem (Theorem 2.2.4) that k-NN with a sequence of random norms in $\mathcal{N}$ (with the random norms being independent of the sample and query) is universally consistent. ∎

With this result, we can take any sequence of random norms, chosen independently of the sample and the query, from a family of norms $\mathcal{N}$ that satisfies certain conditions, and the k-NN learning rule with the resulting sequence of norms is universally consistent. We now provide some examples of universally consistent learning rules based on this theorem.

Corollary 3.2.6. *Let $\mathcal{N}$ be the family of ℓ^p-norms on $\mathbb{R}^d$, with $1 \leq p \leq \infty$. Then the k-NN learning rule, with the norm $\rho_n \in \mathcal{N}$ chosen at each step independently of the sample and the query, is universally consistent.*

Proof: By Lemma 3.1.2, for any ℓ^p norm ρ, $\|\cdot\|_\infty \preceq \rho \preceq \|\cdot\|_1$, hence by Theorem 3.2.5, k-NN with any sequence of random ℓ^p norms is universally consistent (with p being independent of the sample and the query). ∎

For an application of this result, suppose we have a labelled sample (X_1, Y_1), $(X_2, Y_2), \ldots, (X_{2n}, Y_{2n})$ of size $2n$. We can split this sample into two samples (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n)$ and $(X_{n+1}, Y_{n+1}), (X_{n+2}, Y_{n+2}), \ldots, (X_{2n}, Y_{2n})$, which we see are independent of each other. We can then use one of these samples to find a norm (so we

start with a family of norms satisfying the conditions of Theorem 3.2.5 and optimize over norms in the family for this sample), and then use the norm we found as the norm for k-NN with the other sample to classify the query.

3.3 Matrix-based Norms

We now investigate the universal consistency of k-NN when we select a matrix from a family of matrices (based on the dataset), multiply by the matrix, and then apply an existing norm from a family of norms.

The *general linear group* $GL(n)$ is the group of all invertible $n \times n$ matrices. We will now show that multiplication by matrices in $GL(n)$ can be used to create new norms.

Lemma 3.3.1. *Let A be an $n \times n$ invertible matrix (equivalently, $A \in GL(n)$). Then for any norm $\|\cdot\|$ on $\mathbb{R}^d$, $\rho(\boldsymbol{v}) = \|A\boldsymbol{v}\|$ is also a norm.*

Proof: We show that $\rho(\boldsymbol{v}) = \|A\boldsymbol{v}\|$ for $\boldsymbol{v} \in \mathbb{R}^d$ satisfies the conditions for a norm in $\mathbb{R}^d$:

(i) The ρ-norm of the zero vector is zero,

$$\rho(\mathbf{0}) = \|A\mathbf{0}\| = \|\mathbf{0}\| = 0.$$

For any nonzero vector $\boldsymbol{v}$, since A is invertible $A\boldsymbol{v}$ is nonzero, so the $\|A\boldsymbol{v}\|$ norm of $A\boldsymbol{v}$ is strictly positive, so the ρ-norm of $\boldsymbol{v}$ is strictly positive,

$$\rho(\mathbf{v}) = \|A\boldsymbol{v}\| > 0 \text{ since } A\boldsymbol{v} \neq \mathbf{0}.$$

(ii) If we multiply a vector $\boldsymbol{v} \in \mathbb{R}^d$ by a constant $\lambda \in \mathbb{R}$, we see that

$$\begin{aligned} \rho(\lambda\boldsymbol{v}) &= \|A\lambda\boldsymbol{v}\| \\ &= \|\lambda A\boldsymbol{v}\| \end{aligned}$$

$$= |\lambda| \, \|A\boldsymbol{v}\|$$
$$= |\lambda| \, \rho(\boldsymbol{v}).$$

(iii) The triangle inequality holds for the l-norm, for any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ we see that

$$\begin{aligned}\rho(\boldsymbol{u}+\boldsymbol{v}) &= \|A(\boldsymbol{u}+\boldsymbol{v})\| \\ &= \|A\boldsymbol{u} + A\boldsymbol{v}\| \\ &\leq \|A\boldsymbol{u}\| + \|A\boldsymbol{v}\| \\ &= \rho(\boldsymbol{u}) + \rho(\boldsymbol{v}).\end{aligned}$$

∎

Lemma 3.3.2. *Let $\mathcal{M}$ be a family of invertible d by d matrices and $\mathcal{N}$ be a family of norms on $\mathbb{R}^d$, such that there exists a constant $B \geq 1$, such that for all $\boldsymbol{v} \in \mathbb{R}^d$ and $\rho \in \mathcal{N}$, $\frac{1}{B}\|\boldsymbol{v}\| \leq \rho(\boldsymbol{v}) \leq B\|\boldsymbol{v}\|$. If there exists a constant $C \geq 1$ such that $\frac{1}{C}\|\boldsymbol{v}\| \leq \|A\boldsymbol{v}\| \leq C\|\boldsymbol{v}\|$ $\forall \boldsymbol{v} \in \mathbb{R}^d$ $\forall A \in M_{d,d}(\mathbb{R})$, then the family of norms $\mathcal{N}^* = \{\rho^* : \rho^*(\boldsymbol{v}) = \rho(A\boldsymbol{v})$ $\forall A \in M_{d,d}(\mathbb{R})\}$ satisfies the property that there exists some constant $A \geq 1$ such that for all $\rho \in \mathcal{N}^*$ and $\boldsymbol{v} \in \mathbb{R}^d$, $\frac{1}{A}\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\| \leq A\|\boldsymbol{v}\|$.*

Proof: By assumption, there exists a constant $B \geq 1$ such that for all $\boldsymbol{v} \in \mathbb{R}^d$ and $\rho \in \mathcal{N}$,

$$\frac{1}{B}\|\boldsymbol{v}\| \leq \rho(\boldsymbol{v}) \leq B\|\boldsymbol{v}\|.$$

Suppose we have a norm $\rho^* \in \mathcal{N}^*$, which corresponds to the matrix $A \in M_{d,d}(\mathbb{R})$ and norm $\rho \in \mathcal{N}$. We find that

$$\rho^*(\boldsymbol{v}) = \rho(A\boldsymbol{v}) \leq B\|A\boldsymbol{v}\| \leq BC\|\boldsymbol{v}\|$$

and that

$$\rho^*(\boldsymbol{v}) = \rho(A\boldsymbol{v}) \geq \frac{1}{B}\|A\boldsymbol{v}\| \geq \frac{1}{BC}\|\boldsymbol{v}\|.$$

Hence we find that

$$\frac{1}{BC}\|\boldsymbol{v}\| \leq \rho^*(\boldsymbol{v}) \leq BC\|\boldsymbol{v}\|.$$

We therefore find that for $A = BC$, for all $\rho \in \mathcal{N}^*$ and $\boldsymbol{v} \in \mathbb{R}^d$, $\frac{1}{A}\|\boldsymbol{v}\| \leq \|\boldsymbol{v}\| \leq A\|\boldsymbol{v}\|$.

∎

Corollary 3.3.3. *Suppose we apply the k-NN classifier with a sequence of random norms (chosen independently of the sample and the query) from the family of norms $\mathcal{N}^*$ of the form in Lemma 3.3.2. Then the resulting classifier is universally consistent.*

Proof: We have that $\mathcal{N}^*$ is bounded both below and above by Lemma 3.3.2, so it satisfies the conditions in Theorem 3.2.5, and so k-NN with $\mathcal{N}^*$ is universally consistent. ∎

One important family of matrices that we will use is the group of *orthogonal* d by d matrices $O(d)$ (and the subgroup of *special orthogonal* matrices $SO(d)$).

Definition 3.3.4. An d by d matrix Q is an *orthogonal* matrix if multiplication of the matrix by its transpose (in either order) results in the identity matrix:

$$Q^\intercal Q = QQ^\intercal = I_m \tag{3.10}$$

We say that Q is a *special orthogonal* matrix if it satisfies the additional criterion that its determinant is one:

$$\det(Q) = 1 \tag{3.11}$$

An important result is that multiplication by orthogonal matrices preserves the Euclidean norm of a vector.

Theorem 3.3.5. *For any matrix $Q \in O(d)$ and vector $\boldsymbol{x} \in \mathbb{R}^d$, the Euclidean norm of $\boldsymbol{x}$ is equal to the Euclidean norm of $Q\boldsymbol{x}$, $\|\boldsymbol{x}\|_2 = \|Q\boldsymbol{x}\|_2$.*

Proof: This is a standard result, it is proven in Theorem A.1.3 (in Appendix 1) in [28]. ∎

With this result, we are now able show that k-NN with a sequence of random orthogonal matrices $O(d)$ (independent of the sample and the query) is universally consistent.

Corollary 3.3.6. *Suppose we have the family of norms consisting of multiplying the input by a random orthogonal matrix followed by applying an ℓ^p-norm. Then k-NN with a sequence of random norms (independent of the sample and the query) from this family is universally consistent.*

Proof: We see that by Lemma 3.1.2 and Theorem 3.3.5 that the conditions of Lemma 3.3.2 are satisfied for this family of norms. Hence by Corollary 3.3.3, k-NN with this family of norms is universally consistent. ∎

Lemma 3.3.7. *Given $0 < \beta \leq \alpha < \infty$, let $\mathcal{M}_{\beta,\alpha}$ be the family of all diagonal matrices such that for each entry $a_{i,i}$ on the diagonal (with $1 \leq i \leq d$), $\beta \leq |a_{i,i}| \leq \alpha$. Then for any vector $\boldsymbol{x} \in \mathbb{R}^d$ and matrix $D \in \mathcal{M}_{\beta,\alpha}$, $\beta\|\boldsymbol{x}\|_p \leq \|D\boldsymbol{x}\|_p \leq \alpha\|\boldsymbol{x}\|_p$, for any ℓ^p norm.*

Proof: If p is finite, we see that, for any diagonal matrix $D \in \mathcal{M}_{\beta,\alpha}$ with diagonal entries in $[\beta, \alpha]$ and vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$,

$$\begin{aligned}\|A\boldsymbol{x}\|_p &= \sqrt[p]{\sum_{i=1}^{d} (a_i x_i)^p} \\ &\leq \sqrt[p]{\sum_{i=1}^{d} (\alpha x_i)^p}\end{aligned}$$

$$= \alpha \sqrt[p]{\sum_{i=1}^{d} {x_i}^p}$$
$$= \alpha \|\boldsymbol{x}\|_p$$

and that

$$\|A\boldsymbol{x}\|_p = \sqrt[p]{\sum_{i=1}^{d} (a_i x_i)^p}$$
$$\geq \sqrt[p]{\sum_{i=1}^{d} (\beta x_i)^p}$$
$$= \beta \sqrt[p]{\sum_{i=1}^{d} {x_i}^p}$$
$$= \beta \|\boldsymbol{x}\|_p.$$

If p is infinite, we have (with D, $\boldsymbol{x}$ as above)

$$\|A\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq d} \{|a_i x_i|\}$$
$$\leq \max_{1 \leq i \leq d} \{|\alpha x_i|\}$$
$$= \alpha \max_{1 \leq i \leq d} \{|x_i|\}$$
$$= \alpha \|\boldsymbol{x}\|_\infty$$

and that

$$\|A\boldsymbol{x}\|_\infty = \max_{1 \leq i \leq d} \{|a_i x_i|\}$$
$$\geq \max_{1 \leq i \leq d} \{|\beta x_i|\}$$
$$= \beta \max_{1 \leq i \leq d} \{|x_i|\}$$
$$= \beta \|\boldsymbol{x}\|_\infty.$$

∎

Lemma 3.3.8. *Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two families of invertible d by d matrices that both satisfy the following boundedness condition: there exists $b > 0$ such that for all $\boldsymbol{v} \in \mathbb{R}^d$ and $B \in \mathcal{M}_1$, $\frac{1}{b}\|\boldsymbol{v}\| \leq \|B\boldsymbol{v}\| \leq b\|\boldsymbol{v}\|$ (and a corresponding constant $c > 0$ exists for $\mathcal{M}_2$). We define $\mathcal{M}$ to be the product of all pairs of matrices in $\mathcal{M}_1$ and $\mathcal{M}_2$, that is $A \in \mathcal{M}$ if and only if $A = BC$ with $B \in \mathcal{M}_1$ and $C \in \mathcal{M}_2$. Then $\mathcal{M}$ is a bounded family of invertible matrices.*

Proof: We first notice that the product of invertible matrices is invertible. For any matrix $A \in \mathcal{M}$, we let $B \in \mathcal{M}_1$ and $C \in \mathcal{M}_2$ be such that $A = BC$, and $b, c \geq 1$ be constants such that $\frac{1}{b}\|\boldsymbol{v}\| \leq \|B\boldsymbol{v}\| \leq b\|\boldsymbol{v}\|$ for all $\boldsymbol{v} \in \mathbb{R}^d$ and for all $B \in \mathcal{M}_1$, and similarly for C. We notice that for every vector $\boldsymbol{v} \in \mathbb{R}^d$,

$$\begin{aligned} \|A\boldsymbol{v}\| &= \|BC\boldsymbol{v}\| \\ &\geq \frac{1}{b}\|C\boldsymbol{v}\| \\ &\geq \frac{1}{bc}\|\boldsymbol{v}\| \end{aligned}$$

and that

$$\begin{aligned} \|A\boldsymbol{v}\| &= \|BC\boldsymbol{v}\| \\ &\leq b\|C\boldsymbol{v}\| \\ &\leq bc\|\boldsymbol{v}\|. \end{aligned}$$

∎

From this result we see that we can, for instance, first multiply by a diagonal matrix from a bounded family and then multiply by an orthogonal matrix (or in the other order), and retain universal consistency.

We would now like to find a general criterion for checking if a family of matrices is bounded both below and above. We can do this from the *singular value decomposition* of a matrix.

Theorem 3.3.9. *Let A be a real valued d by d matrix. There exist orthogonal d by d matrices U and V and an d by d diagonal matrix Σ such that the diagonal entries of Σ are the square roots of the eigenvalues of $A^\intercal A$ and of $AA^\intercal$ (they are called the* singular values *of A) and*

$$A = U\Sigma V^\intercal. \tag{3.12}$$

We call this the singular value decomposition *of the matrix A.* [1]

Proof: This is a standard result, a proof can be found in [5] (Chapter 8, Theorem 8.19). ∎

Lemma 3.3.10. *For any vector $\boldsymbol{v} \in \mathbb{R}^d$ and d by d diagonal matrix D, if we let $a_1, a_2, \ldots, a_d$ be the entries on the diagonal of D and $\boldsymbol{v} = (v_1, v_2, \ldots, v_d)$, then*

$$\min_{1\leq i\leq d} |a_i| \, \|\boldsymbol{v}\|_2 \leq \|D\boldsymbol{v}\|_2 \leq \max_{1\leq i\leq d} |a_i| \, \|\boldsymbol{v}\|_2. \tag{3.13}$$

Proof: We notice that for the second inequality,

$$\begin{aligned}
\|D\boldsymbol{v}\|_2 &= \sqrt{(a_1v_1)^2 + (a_2v_2)^2 + \cdots + (a_dv_d)^2} \\
&= \sqrt{a_1^2v_1^2 + a_2^2v_2^2 + \cdots + a_d^2v_d^2} \\
&\leq \sqrt{\left(\max_{1\leq i\leq d} a_i^2\right)(v_1^2 + v_2^2 + \cdots + v_d^2)} \\
&= \left(\max_{1\leq i\leq d} |a_i|\right)\sqrt{(v_1^2 + v_2^2 + \cdots + v_d^2)} \\
&= \left(\max_{1\leq i\leq d} |a_i|\right)\|\boldsymbol{v}\|_2.
\end{aligned}$$

[1] A version of this result also holds for complex matrices and non-square matrices, for our purposes the result for square real-valued matrices suffices.

Similarly, for the first inequality, we have

$$\begin{aligned}
\|D\boldsymbol{v}\|_2 &= \sqrt{(a_1 v_1)^2 + (a_2 v_2)^2 + \cdots + (a_d v_d)^2} \\
&= \sqrt{a_1^2 v_1^2 + a_2^2 v_2^2 + \cdots + a_d^2 v_d^2} \\
&\geq \sqrt{\left(\min_{1\leq i\leq d} a_i^2\right)(v_1^2 + v_2^2 + \cdots + v_d^2)} \\
&= \left(\min_{1\leq i\leq d} |a_i|\right)\sqrt{(v_1^2 + v_2^2 + \cdots + v_d^2)} \\
&= \left(\min_{1\leq i\leq d} |a_i|\right)\|\boldsymbol{v}\|_2.
\end{aligned}$$

∎

Theorem 3.3.11. *Let A be a real-valued d by d matrix and $A = U\Sigma V^\intercal$ be its singular value decomposition, with $\sigma_1, \sigma_2, \ldots, \sigma_m$ being the singular values of A. We have that*

$$\left(\min_{1\leq i\leq d} |\sigma_i|\right)\|\boldsymbol{v}\|_2 \leq \|A\boldsymbol{v}\|_2 \leq \left(\max_{1\leq i\leq d} |\sigma_i|\right)\|\boldsymbol{v}\|_2. \tag{3.14}$$

Proof: For the upper bound inequality, we see that

$$\begin{aligned}
\|A\boldsymbol{v}\| &= \|U\Sigma V^\intercal \boldsymbol{v}\|_2 \\
&= \|\Sigma V^\intercal \boldsymbol{v}\|_2 \text{ since } U \text{ is orthogonal} \\
&\leq \left(\max_{1\leq i\leq d} |\sigma_i|\right)\|V^\intercal \boldsymbol{v}\|_2 \text{ by Lemma 3.3.10} \\
&= \left(\max_{1\leq i\leq d} |\sigma_i|\right)\|\boldsymbol{v}\|_2 \text{ since } V^\intercal \text{ is orthogonal.}
\end{aligned}$$

Similarly for the lower bound inequality, we see that

$$\begin{aligned}
\|A\boldsymbol{v}\| &= \|U\Sigma V^\intercal \boldsymbol{v}\|_2 \\
&= \|\Sigma V^\intercal \boldsymbol{v}\|_2 \text{ since } U \text{ is orthogonal} \\
&\geq \left(\min_{1\leq i\leq d} |\sigma_i|\right)\|V^\intercal \boldsymbol{v}\|_2 \text{ by Lemma 3.3.10}
\end{aligned}$$

$$= \left(\min_{1 \leq i \leq d} |\sigma_i| \right) \|\boldsymbol{v}\|_2 \text{ since } V^\intercal \text{ is orthogonal.}$$

∎

From this, we see that we can multiply the data by a matrix from a family of matrices whose singular values are bounded (both above and from below away from zero) and maintain universal consistency.

Corollary 3.3.12. *If $\mathcal{N}$ is a bounded family of norms and $\mathcal{M}$ is a family of matrices whose singular values are bounded below away from zero and are bounded above by some finite value, then k-NN with a sequence of random norms (independent of the sample and the query) from the family of norms consisting of first multiplying by a matrix in $\mathcal{M}$ and then applying a norm in $\mathcal{N}$ is universally consistent.*

Proof: This follows from Theorem 3.3.11 and Corollary 3.3.3. ∎

3.4 Sequences of Norms that Depend on the Sample

In Theorem 3.2.5, we have assumed that the sequence of norms is independent of the sample and the query. This is a strong assumption we would like to eliminate. Unfortunately, removing this assumption appears to be quite difficult.

In [14], Theorem 26.3, it is claimed that a result of a form similar to Theorem 3.2.5 holds. The book claims that if we multiply the data by a matrix A_n that is a function of the sample points $X_1, X_2, \ldots, X_n$ and then apply the Euclidean norm, then k-NN with the resulting distance is universally consistent. The proof is performed by checking the three conditions of Stone's theorem. Unfortunately, the proof that

such a classifier satisfies the first condition of Stone's theorem is incorrect. The book makes the following argument: for the first condition we need that the number of data points that can be among the k nearest neighbours of X is at most kc_d, where c_d is a constant that depends on the dimension only; and that this is a deterministic property that can be proven in exactly the same manner as for the usual k-NN learning rule.

The problem with this argument is that if the norm is a function of the sample, when we do the exchange of random variables in the proof of Stone's lemma (Lemma 2.3.8), we obtain a different norm for each point, as shown below:

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[W_{ni}(X) f(X_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{k} \mathbb{1}_{\{X_i \text{ is a } k\text{-nearest neighbour of } X \text{ in the } \rho_n(X_1,X_2,\ldots,X_n) \text{ norm among } X_1,X_2,\ldots,X_n\}} f(X_i)\right] \\
&= \frac{1}{k}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\left\{\substack{X \text{ is a } k\text{-nearest neighbour of } X_i \text{ in the } \rho_n(X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n) \\ \text{norm among } X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n}\right\}} f(X)\right]
\end{aligned}$$

We see that we have a different norm for each point, namely $\rho_n(X, X_2, X_3, \ldots, X_n)$ for the point X_1, $\rho_n(X_1, X, X_3, \ldots, X_n)$ for the point X_2, and so on. This means we must consider the set of points such that X can be the k-nearest neighbour of them in any of the norms from the family, and not just each norm by itself. As we show in the following example, the geometrical argument used in Stone's lemma does not work if we require a bound on the number of points that can be considered a nearest neighbour for *any* norm in a family (even if the family of norms is bounded, as described in Theorem 3.2.5) as opposed to a single norm.

Example 3.4.1. Suppose the query X is at the origin, and the data points $X_1, X_2, \ldots, X_n$ are arranged on the upper-right part of the unit circle around X, as shown in Figure 3.1. We can accomplish this by taking the coordinate of the point X_i to be as follows:

$$X_i = \left(\cos\left(\frac{(i-1)\pi}{2n}\right), \sin\left(\frac{(i-1)\pi}{2n}\right)\right), \ 1 \leq i \leq n \tag{3.15}$$

We then define a norm ρ_i for each $1 \leq i \leq n$ as follows: we first apply a rotation of angle $-(i-1)\pi/(2n)$ (effectively rotating the unit circle such that the point X_i

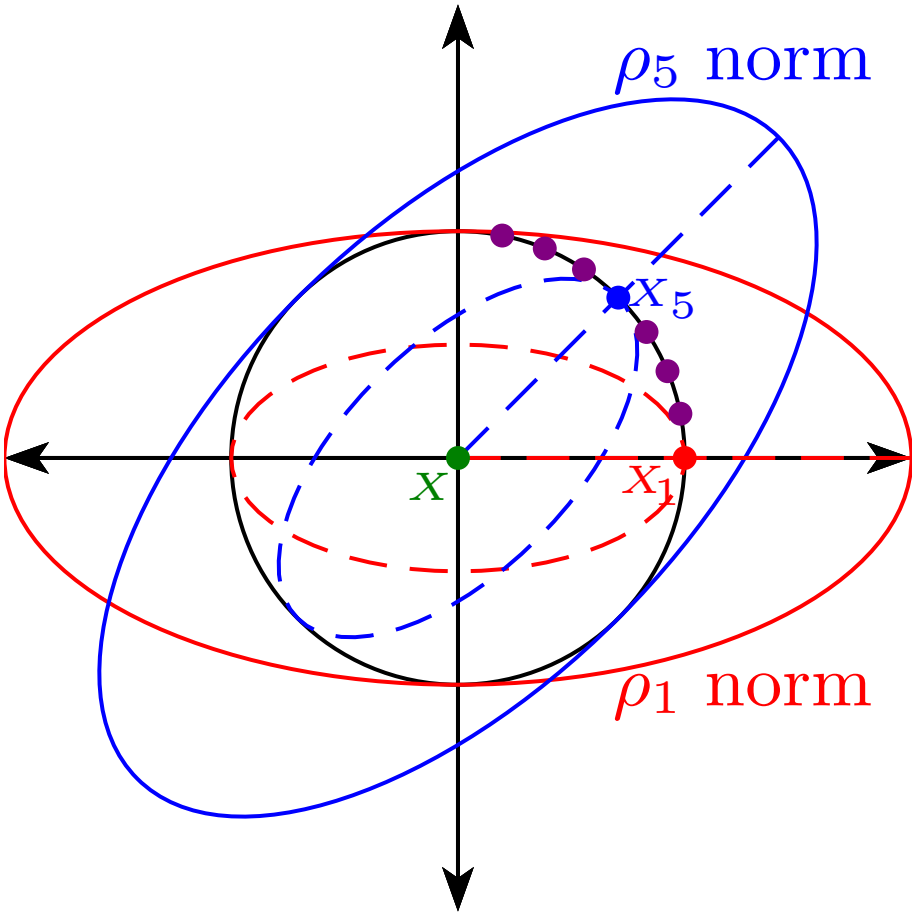

Figure 3.1: We see that for each of the points $X_1, X_2, \ldots, X_n$ (with $n = 8$ in this illustration), we have that X_i is the nearest neighbour to the origin X in the ρ_i norm. The points $X_1, X_2, \ldots, X_n$ are distributed along the circle as described by equation (3.15) (with corresponding norms given by (3.16)).

is now at $(1,0)$), we then multiply by the matrix $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ and apply the Euclidean norm. This norm is given by the formula

$$\begin{aligned} \rho_i((x,y)) &= \left\| \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \cos\left(\frac{(i-1)\pi}{2n}\right) & \sin\left(\frac{(i-1)\pi}{2n}\right) \\ -\sin\left(\frac{(i-1)\pi}{2n}\right) & \cos\left(\frac{(i-1)\pi}{2n}\right) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right\|_2 \\ &= \left\| \begin{bmatrix} \cos\left(\frac{(i-1)\pi}{2n}\right) & \sin\left(\frac{(i-1)\pi}{2n}\right) \\ -2\sin\left(\frac{(i-1)\pi}{2n}\right) & 2\cos\left(\frac{(i-1)\pi}{2n}\right) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \right\|_2 \\ &= \sqrt{\left(3\sin^2\left(\frac{(i-1)\pi}{2n}\right)+1\right)x + \left(3\cos^2\left(\frac{(i-1)\pi}{2n}\right)+1\right)y}. \end{aligned} \tag{3.16}$$

This norm ρ_i is intuitively a norm that gives twice as much importance to the y-axis as to the x-axis, with the axes rotated by angle $(i-1)\pi/(2n)$ prior to applying the norm (without rotating the points). We easily see that X_i is the nearest point to the origin X among $X_1, X_2, \ldots, X_n$ in the ρ_i norm. This holds for each of the norms in $\rho_1, \rho_2, \ldots, \rho_n$ and the corresponding data point, hence we see that there are n points in the sample that can be the nearest neighbour of the query X for some norm $\rho_1, \rho_2, \ldots, \rho_n$ (with a different norm for each point). In Stone's lemma (Lemma 2.3.8), we need that the number of such points is at most ck, with c being a fixed constant. Since there are n such points, if we substitute this into the inequality (instead of ck) we obtain an upper bound of $\frac{n}{k}E\left[f(X)\right]$, which is not useful for us as $n/k \to \infty$ as $n \to \infty$ (since $k/n \to 0$ as $n \to \infty$). Hence we see that even though the argument in Stone's lemma works for each fixed norm among $\rho_1, \rho_2, \ldots, \rho_n$, it does not work for the combined family of all such norms.

The family of norms of the form in equation (3.16) (containing all such norms for any $n \geq 1$ and $1 \leq i \leq n$) is bounded (in the sense of Theorem 3.2.5), since the rotation matrix is an orthogonal matrix and all entries of the fixed diagonal matrix are nonzero, so (by Theorem 3.3.5 and Lemmas 3.3.7 and 3.3.8) the family of norms consisting of first applying the rotation, then multiplying by the diagonal matrix and

then applying the Euclidean norms is bounded both above and below (that is, for any norm ρ in this family, there exists $C \geq 1$ such that $\frac{1}{C}\|\boldsymbol{v}\|_2 \leq \rho(\boldsymbol{v}) < C\|\boldsymbol{v}\|$). Therefore, this family of norms cannot be excluded by the boundedness conditions on the family of norms, since it satisfies these conditions of the theorem.

This example above shows that our proof for universal consistency will not work if we replace the independently chosen norm by a norm chosen as some function of the sample data points. To prove universal consistency with the norm as a function of the sample (and possibly the query), we will need to use additional or different techniques. A possible approach would be to show that the probability of a configuration of points such as the one described above occurs with a probability that decreases sufficiently fast as n approaches infinity, for every probability measure on $\mathbb{R}^d$ (the deterministic geometric result would then be replaced by a probabilistic argument).

3.5 Necessity of the Boundedness Conditions

Our theorem requires that there exist norms $\|\cdot\|_L$ and $\|\cdot\|_U$ such that for any norm ρ in our family of norms, $\|\cdot\|_L \preceq \rho \preceq \|\cdot\|_U$. We now see that both the conditions that the family of norms is bounded from above and from below are necessary.

Suppose we have the probability measure μ on $\mathbb{R}^2 \times \{0,1\}$, such that for any $A \subseteq \mathbb{R}^2 \times \{0,1\}$ (where λ is the Lebesgue measure on $\mathbb{R}$):

$$\mu(A) = \frac{1}{2}\lambda(\{x : (x,0) \times \{0\} \in A, x \in [0,1]\}) + \frac{1}{2}\lambda(\{x : (x,1) \times \{1\} \in A, x \in [0,1]\}) \tag{3.17}$$

That is, there is a line segment $\{(x,0) \ : x \in [0,1]\}$ with uniform probability density $1/2$ at $y = 0$ with label 0, and a line segment $\{(x,1) \ : x \in [0,1]\}$ with uniform probability density $1/2$ at $y = 1$ with label 1. Let X be a query and $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n iid sample points. We define $X_{i,x}$ to be the x coordinate of the point and $X_{i,y}$ to be the y coordinate of the point X_i (for the point X, we

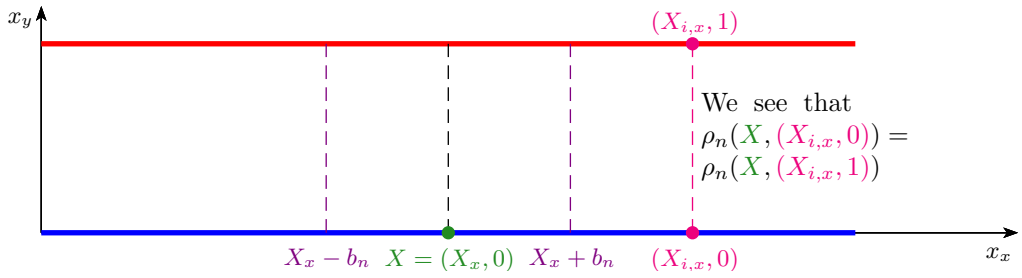

Figure 3.2: If we have a point X_i whose x-coordinate differs by more than b_n from X, then the ρ_n-distance of X_i from X is $\rho_n(X, X_i) = a_n |X_{i,x} - X_x|$, and does not depend whether the point is on the upper or the lower line segment.

define X_x to be the x coordinate and X_y to be the y coordinate). We also define $X_{(i)}$ to be the i^{th} point from $X_1, X_2, \ldots, X_n$ in distance from X, in a given norm.

Suppose we have a sequence of norms $(\rho_n)_{n=1}^{\infty}$ of the form (with $(a_n)_{n=1}^{\infty}$, $(b_n)_{n=1}^{\infty}$ being two numeric sequences):

$$\rho_n(\boldsymbol{v}) = \left\| \begin{pmatrix} a_n & 0 \\ 0 & b_n \end{pmatrix} \boldsymbol{v} \right\|_{\infty} \tag{3.18}$$

The distance between the query X and a point X_i is

$$\rho_n(X, X_i) = \max\left\{a_n |X_{i,x} - X_x|, b_n |X_{i,y} - X_y|\right\}. \tag{3.19}$$

Suppose that the distance between X and X_i is strictly greater than b_n. Since $b_n |X_{i,y} - X_y| \leq b_n$, it follows that $a_n |X_{i,x} - X_x|$ is the larger term, and so

$$\rho_n(X, X_i) = a_n |X_{i,x} - X_x| \text{ if } \rho_n(X, X_i) > b_n. \tag{3.20}$$

We notice that the strict inequality $\rho_n(X, X_i) > b_n$ holds if and only if $X_{i,x} \in [0, 1] \setminus [X_x - b_n/a_n, X_x + b_n/a_n]$ (no matter what $Y_i = X_{i,y}$ is). This means we have equal length intervals of equal probability density on which the condition holds, and so we find that conditioning on this results in a probability of $1/2$ of the point being

on either segment (and hence a 1/2 probability of each label),

$$\begin{aligned}\mathbb{P}\left(Y_i = 0 | \rho_n(X, X_i) > b_n\right) &= \mathbb{P}\left(Y_i = 1 | \rho_n(X, X_i) > b_n\right) \\ &= \frac{1}{2}.\end{aligned} \tag{3.21}$$

If all the points $X_1, X_2, \ldots, X_n$ satisfy the property that $\rho_n(X, X_i) > b_n$ (or equivalently, the condition $\rho_n(X, X_{(1)}) > b_n$ holds), then by the formula 3.20 we see that the distance of any of the points $X_1, X_2, \ldots, X_n$ from X does not depend on $Y_1, Y_2, \ldots, Y_n$, and by equation 3.21 we have that the points are equally likely to be on either segment. We easily see that the order statistics $(X_{(1)}, Y_{(1)}), (X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ therefore do not depend on which segments the points are on, as long as $\rho_n(X, X_i) > b_n$ holds for each point. From this we see that the order statistics of the points are conditionally independent of the Y_i, if $\rho_n(X, X_{(1)}) > b_n$ holds. Indeed, all the order statistics are equally likely to be on either line segment, as long as the $\rho_n(X, X_i) > b_n$ condition holds, that is for all $i \in \{1, 2, \ldots, n\}$,

$$\begin{aligned}&\mathbb{P}\left(Y_{(i)} = 0 \middle| \forall j \in \{1, 2, \ldots, n\}\ \rho_n(X, X_j) > b_n\right) \\ &= \mathbb{P}\left(Y_{(i)} = 1 \middle| \forall j \in \{1, 2, \ldots, n\}\ \rho_n(X, X_j) > b_n\right) \\ &= \frac{1}{2}.\end{aligned} \tag{3.22}$$

In addition, since the Y_i are iid and are conditionally independent of the distance of the i^{th} order statistic from the query X (as long as $\rho_n(X, X_{(1)}) > b_n$), we see that the $Y_{(i)}$ are conditionally independent of each other. Hence, if $\rho_n(X, X_{(1)}) > b_n$, then the $Y_{(i)}$ are iid Bernoulli random variables with probability 1/2 of being 1 and 1/2 of being 0.

We now need to find a lower bound for the probability that for all $i \in \{1, 2, \ldots, n\}$, $\rho_n(X, X_i) > b_n$. We find that

$$\mathbb{P}(\rho_n(X, X_1) > b_n \text{ and } \ldots \text{ and } \rho_n(X, X_n) > b_n)$$

$$
\begin{aligned}
&= \prod_{i=1}^{n} \mathbb{P}(\rho_n(X, X_i) > b_n) \\
&= (\mathbb{P}(\rho_n(X, X_1) > b_n))^n \\
&= \left(\mathbb{P}\left(X_{1,x} \in [0,1] \setminus \left[X_x - \frac{b_n}{a_n}, X_x + \frac{b_n}{a_n} \right] \right) \right)^n \\
&= \left(\mathbb{E}\left[\mathbb{P}\left(X_{1,x} \in [0,1] \setminus \left[X_x - \frac{b_n}{a_n}, X_x + \frac{b_n}{a_n} \right] \middle| X_x \right) \right] \right)^n \\
&= \left(\mathbb{E}\left[\mu\left(\left([0,1] \setminus \left[X_x - \frac{b_n}{a_n}, X_x + \frac{b_n}{a_n} \right] \right) \times \{0,1\} \right) \right] \right)^n \\
&\geq \left(1 - \frac{2b_n}{a_n} \right)^n.
\end{aligned}
$$

The resulting bound is

$$
\mathbb{P}(\rho_n(X, X_1) > b_n \text{ and } \ldots \text{ and } \rho_n(X, X_n) > b_n) \geq \left(1 - \frac{2b_n}{a_n} \right)^n. \tag{3.23}
$$

Equivalently, we see that probability that the closest order statistic is closer than b_n to the query is bounded from below by $\left(1 - \frac{2b_n}{a_n}\right)^n$.

Let k be any odd integer between 1 and n, $1 \leq k \leq n$. Suppose the query X is on the lower axes, with $Y = 0$ (this occurs with probability 1/2). We then apply k-NN with ρ_n as the choice of norm. If $\rho_n(X, X_i) > b_n$ holds for all $1 \leq i \leq n$, then as we have seen above, the $Y_{(i)}$ are iid Bernoulli random variables with 1/2 probability of being one. Hence the distribution of the fraction of the k nearest points to X that take value one is $C = \frac{1}{k}\text{Binomial}(1/2, k)$. We find that for all odd k, $\mathbb{P}(C \geq 1/2) = 1/2$ (we suppose that k is odd to avoid the case of ties, which slightly complicates our discussion). By symmetry, the argument holds the same way if the query X is on the upper axes, hence the conditional probability of a point being misclassified if $\rho_n(X, X_i) > b_n$ is 1/2.

Now, suppose we have the sequence of norms that is unbounded above (of the form in equation (3.18) with $a_n = n^2$, $b_n = 1$):

$$
\rho_n(\boldsymbol{v}) = \left\| \begin{pmatrix} n^2 & 0 \\ 0 & 1 \end{pmatrix} \boldsymbol{v} \right\|_\infty \tag{3.24}
$$

We then have, by equation (3.23), that

$$\begin{aligned}\mathbb{P}(\rho_n(X, X_1) > 1 \text{ and } \ldots \text{ and } \rho_n(X, X_n) > 1) &\geq \left(1 - \frac{2}{n^2}\right)^n \\ &= \frac{(n^2-2)^n}{n^{2n}} \\ &= \frac{n^{2n} + \mathcal{O}(n^{2n-1})}{n^{2n}} \to 1 \text{ as } n \to \infty.\end{aligned}$$

This means that the probability of all of the sample points being at least distance 1 from the query approaches 1, and hence we see that the probability of a query being misclassified approaches $1/2$ as $n \to \infty$. Since the Bayes error is zero here, this means that k-NN with this sequence of norms is not consistent with this distribution.

Now, suppose we have the sequence of norms that is not bounded from below by any norm (of the form in equation (3.18) with $a_n = 1$, $b_n = n^2$):

$$\rho_n(\boldsymbol{v}) = \left\| \begin{pmatrix} 1 & 0 \\ 0 & 1/n^2 \end{pmatrix} \boldsymbol{v} \right\|_\infty \tag{3.25}$$

Similarly to the previous case, we have by equation (3.23) that

$$\begin{aligned}\mathbb{P}(\rho_n(X, X_1) > 1/n^2 \text{ and } \ldots \text{ and } \rho_n(X, X_n) > 1) &\geq \left(1 - \frac{2}{n^2}\right)^n \\ &= \frac{(n^2-2)^n}{n^{2n}} \\ &= \frac{n^{2n} + \mathcal{O}(n^{2n-1})}{n^{2n}} \to 1 \text{ as } n \to \infty.\end{aligned}$$

Similarly to the previous case, we see that the misclassification error approaches $1/2$ as $n \to \infty$, and hence k-NN with this sequence of norms is not consistent with this distribution μ.

We now see that k-NN with the sequences of norms (3.24) and (3.25) is not universally consistent. Both sequences of norms do not satisfy the boundedness conditions required in Theorem 3.2.5. The first sequence (3.24) is unbounded above, since if we take the vector $\boldsymbol{v} = (1, 0)$, then $\rho_n(\boldsymbol{v}) = n^2 \to \infty$ as $n \to \infty$. The second

sequence (3.25) is unbounded below by any norm, since if we take the nonzero vector $\boldsymbol{w} = (0, 1)$, $\rho_n(\boldsymbol{w}) = 1/n^2 \to 0$ as $n \to \infty$, while the norm of any nonzero vector is nonzero. Hence we see that we need to bound the sequence of norms from above and from below, otherwise universal consistency may not hold.

3.6 Failure of the Geometric Stone's Lemma for Quasinorms

We now give an example that shows that the geometric Stone's Lemma does not hold for the ℓ^p quasinorms, with $0 < p < 1$. In particular, we show that it fails for the $\ell^{1/2}$ quasinorm on $\mathbb{R}^2$ (which we denote ρ), that no cone can contain both the vector $(1, 0)$ and vectors nearby with nonzero y-component and satisfy the property that if $\boldsymbol{x}, \boldsymbol{y}$ are vectors in the cone with $0 < \rho(\boldsymbol{x}) < \rho(\boldsymbol{y})$, then $\rho(\boldsymbol{y} - \boldsymbol{x}) < \rho(\boldsymbol{y})$. We further show that this problem cannot be avoided by considering axis vectors separately from vectors that are not on an axis.

Example 3.6.1. Suppose we have the point $\boldsymbol{y} = (1, 0)$ in $\mathbb{R}^2$ and a value $r \in (0, 1)$ such that $(1, r)$ is still in the cone. Any positive multiple of a vector inside the cone is in the cone, so it follows that $\boldsymbol{x} = r(1, r) = (r, r^2)$ is in the cone as well. If we take ρ to be the $\ell^{1/2}$ quasinorm and $0 < r < 1/4$, we find that

$$\begin{aligned}
\rho(\boldsymbol{x}) &= \rho\left(\left(r, r^2\right)\right) \\
&= \left(\sqrt{|r|} + \sqrt{|r^2|}\right)^2 \\
&= r^2 + 2r\sqrt{r} + r \\
&\leq \left(\frac{1}{4}\right)^2 + 2\left(\frac{1}{4}\right)\sqrt{\frac{1}{4}} + \frac{1}{4} \\
&= 0.5625 \\
&< 1 = \rho(\boldsymbol{y}).
\end{aligned}$$

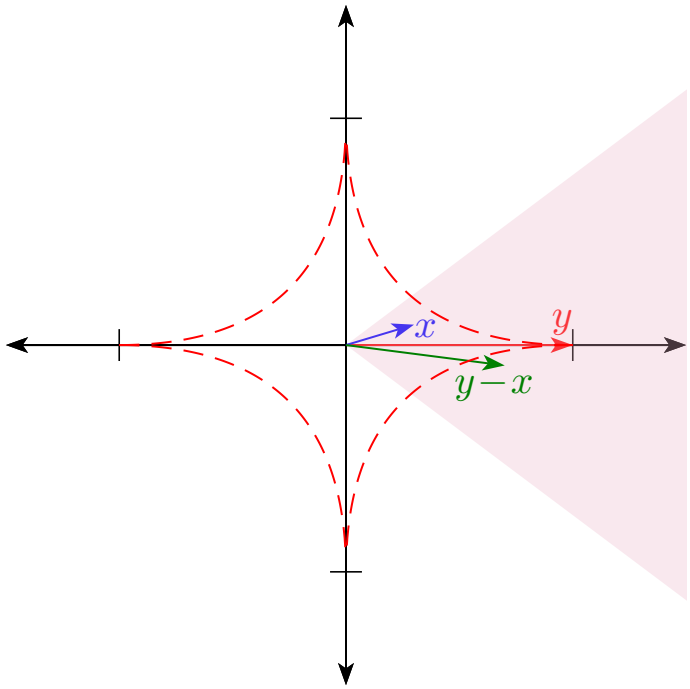

Figure 3.3: In this illustration we have that $\rho(\boldsymbol{y}-\boldsymbol{x}) > \rho(\boldsymbol{x})$ while $0 < \rho(\boldsymbol{x}) < \rho(\boldsymbol{y})$ (with ρ being the $\ell^{1/2}$ quasinorm). In this example $\rho(\boldsymbol{y}) = 1$. The dashed curve is the unit sphere (in the $\ell^{1/2}$ quasinorm). We see that the $\boldsymbol{y}-\boldsymbol{x}$ vector lies outside the unit ball. Such pairs of vectors $\boldsymbol{x}, \boldsymbol{y}$ are possible for arbitrarily thin cones around an axis in the $\ell^{1/2}$ quasinorm.

We additionally find that

$$\begin{aligned}\rho(\boldsymbol{y}-\boldsymbol{x}) &= \rho\left((1,0)-\left(r,r^2\right)\right)\\ &= \rho\left(\left(1-r,-r^2\right)\right)\\ &= \left(\sqrt{|1-r|}+\sqrt{|-r^2|}\right)^2\\ &= 1-r+2r\sqrt{1-r}+r^2.\end{aligned}$$

We now define the function $f(r)$ by subtracting $\rho(\boldsymbol{y})$ from $\rho(\boldsymbol{y}-\boldsymbol{x})$,

$$f(r) = r^2+2r\sqrt{1-r}-r. \tag{3.26}$$

We need to show that $f(r) > 0$ for all $r \in (0, 1/4)$. We find that $f(0) = 0$ and the derivative is

$$f'(r) = 2r+\frac{r}{\sqrt{1-r}}+2\sqrt{1-r}-1. \tag{3.27}$$

We see that for all $0 < r < 1/4$, $2r+\frac{r}{\sqrt{1-r}} \geq 0$, and that

$$2\sqrt{1-r}-1 \geq 2\sqrt{1-\frac{1}{4}}-1 \approx 0.732051.$$

It follows that $f(r) > 0$ for all $r \in (0, 1/4)$, since f is strictly increasing on this interval and $f(0) = 0$. This means that $\rho(\boldsymbol{y}-\boldsymbol{x})-\rho(\boldsymbol{y}) > 0$, or equivalently $\rho(\boldsymbol{y}-\boldsymbol{x}) > \rho(\boldsymbol{y})$. We have earlier found that $\rho(\boldsymbol{x}) < \rho(\boldsymbol{y})$. Hence we see that no cone that satisfies the condition of the geometric Stone's lemma (that if $\boldsymbol{x}, \boldsymbol{y}$ are vectors in the cone with $0 < \rho(\boldsymbol{x}) < \rho(\boldsymbol{y})$, then $\rho(\boldsymbol{y}-\boldsymbol{x}) < \rho(\boldsymbol{y})$) can contain both an axis vector and any vector that does not lie on that axis.

We see that ρ is continuous, which means that if $\rho(\boldsymbol{y}-\boldsymbol{x}) > \rho(\boldsymbol{y})$ and $\rho(\boldsymbol{x}) < \rho(\boldsymbol{y})$, there exists a neighbourhood of radius $\delta > 0$ around $\boldsymbol{y}$ such that for all $\boldsymbol{y}' \in B_r(\boldsymbol{y}, \|\cdot\|)$, $\rho(\boldsymbol{y}'-\boldsymbol{x}) > \rho(\boldsymbol{y}')$ and $\rho(\boldsymbol{x}) < \rho(\boldsymbol{y}')$. It follows that we can replace $\boldsymbol{y}$ with $\boldsymbol{y}' = (1, \delta/2)$, which has a nonzero y-coordinate. From this we see that simply putting axis vectors in their own category does not fix the above problem, any cone that contains vectors arbitrarily close to an axis has this problem.

By this example, we see that we cannot prove the universal consistency of k-NN for quasinorms using the classical approach with Stone's lemma with cones. This does not necessarily mean that k-NN is inconsistent with ℓ^p-quasinorms (indeed, this would be very surprising), it simply means that the geometric Stone's lemma with cones cannot be used to prove this result.

Chapter 4

k-NN with a Sequence of Random Uniformly Locally Lipschitz Functions

We would like to generalize norms. Suppose we have a function $f : \mathbb{R}^d \to \mathbb{R}$, it can serve as a function to measure the "distance" between two points as follows: given points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we compute $f(\boldsymbol{x} - \boldsymbol{y})$ and take this to be our distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. We require f to be nonnegative and that $f(\boldsymbol{x}) = 0$ only at $\boldsymbol{x} = 0$. We do not require that f satisfies the triangle inequality or absolute homogeneity (which norms have to satisfy). In particular, we do not require the distance we have just defined (in terms of f) to be a metric.

We can now consider k-NN with the distance function f (that is, given points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we take $f(\boldsymbol{x} - \boldsymbol{y})$ as the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ for k-NN). Intuitively, any such function that is increasing away from zero is a possible candidate for k-NN. We illustrate some examples of such functions in Figure 4.1. In this section, we show that k-NN is universally consistent with such a function (or more generally, a sequence of such functions independent of the sample and the query) under certain conditions

(in particular, that the family of functions we consider is uniformly locally Lipschitz near zero (with respect to the Euclidean norm, or equivalently any other norm on $\mathbb{R}^d$), in addition to a few other conditions). The question of universal consistency of the k-NN under quasinorms remains open, since quasinorms are not necessarily locally Lipschitz near the origin (in particular, this condition does not hold for ℓ^p quasinorms on $\mathbb{R}^d$ with $0 < p < 1$, even though the ℓ^p quasinorms are uniformly continuous on $\mathbb{R}^d$ with respect to the Euclidean norm).

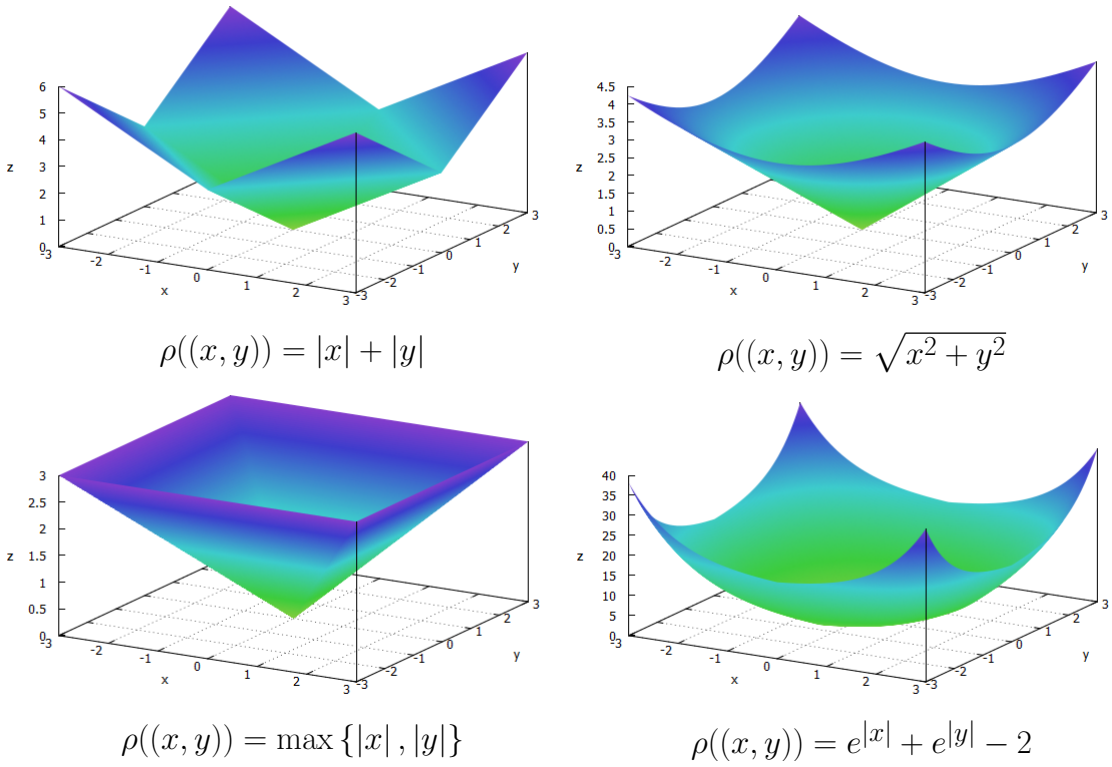

Figure 4.1: Illustration showing the graphs of the ℓ^1 norm (top left), ℓ^2 norm (top right), ℓ^∞ norm (bottom left), and $\rho((x,y)) = e^{|x|} + e^{|y|} - 2$ (bottom right) as functions on $\mathbb{R}^2$. We see that all of these functions are increasing as we move away from the origin, and we can use them as distances with k-NN (using the method we discussed). We would like to establish that k-NN with functions like the one on the bottom right is universally consistent.

4.1 General Theory

Let $\mathcal{F}$ be a family of measurable functions on $(\mathbb{R}^d, \|\cdot\|)$ such that there exist constants $\alpha, \beta, \gamma > 0$ and a radius $r > 0$ where:

1. All $\rho \in F$ are Lipschitz with constant α on the domain $B_r(x)$, that is, for all $\boldsymbol{x}, \boldsymbol{y} \in B_r(x)$,
$$|\rho(\boldsymbol{x}) - \rho(\boldsymbol{y})| \leq \alpha \|\boldsymbol{x} - \boldsymbol{y}\|. \tag{4.1}$$

2. For any nonzero vector $\boldsymbol{v} \in B_r(\mathbf{0}, \|\cdot\|)$ (with $\boldsymbol{v} \neq \mathbf{0}$), if we define $f(\lambda) = \rho(\lambda \boldsymbol{v})$, then $f'(\lambda) \geq \beta \|\boldsymbol{v}\|$ for all $0 < \lambda < 1$. [1]

3. Outside of the ball of radius r, ρ is bounded from below by γ, that is, for all $\boldsymbol{v} \in \mathbb{R}^d$ with $\|\boldsymbol{v}\| \geq r$, $\rho(\boldsymbol{v}) \geq \gamma$.

4. The function ρ is symmetric, $\rho(\boldsymbol{x}) = \rho(-\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$. [2]

5. The function ρ takes value zero at the point zero, $\rho(\mathbf{0}) = 0$.

We may assume without loss of generality that $\alpha \geq 1$ and that $\beta \leq 1$, this simplifies some of our proofs.

We show that k-NN is universally consistent with any sequence of functions from the family $\mathcal{F}$ that is independent of the sample and the query (like Theorem 3.2.5, but with family of norms replaced by a family of uniformly locally Lipschitz functions). We do this by showing that the conditions of Stone's Theorem (Theorem 2.2.4) are satisfied. For a query X at step n, we define $X_{(1,\rho)}, X_{(2,\rho)}, \ldots, X_{(n,\rho)}$ to be the order statistics in increasing distance from X, with ρ as the distance. We define the weight function to be

$$W_{ni}(X) = \begin{cases} \dfrac{1}{k} & \text{if } X_i \in \{X_{(1,\rho)}, X_{(2,\rho)}, \ldots, X_{(k,\rho)}\} \\ 0 & \text{otherwise.} \end{cases} \tag{4.2}$$

[1] It should be possible to replace this condition by a similar lower bound with the limit in the derivative replaced by liminf.

[2] It is almost certainly possible to remove or at least weaken this condition.

Lemma 4.1.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Lipschitz continuous function on $B_r(\mathbf{0}, \|\cdot\|)$ with Lipschitz constant α. Then for all $\boldsymbol{x}, \boldsymbol{y}$ such that $\|\boldsymbol{x}\| \leq r$, $\|\boldsymbol{y}\| \leq r$, and $\|\boldsymbol{x} + \boldsymbol{y}\| \leq r$, we have the inequality*

$$|f(\boldsymbol{x} + \boldsymbol{y})| \leq |f(\boldsymbol{x})| + \alpha\|\boldsymbol{y}\|. \tag{4.3}$$

Proof: We notice that since f has Lipschitz constant α on $B_r(\mathbf{0}, \|\cdot\|)$, for all $\boldsymbol{x}', \boldsymbol{y}' \in B_r(\mathbf{0}, \|\cdot\|)$,

$$|f(\boldsymbol{x}') - f(\boldsymbol{y}')| \leq \alpha\|\boldsymbol{x}' - \boldsymbol{y}'\|. \tag{4.4}$$

By the reverse triangle inequality, we have that $|f(\boldsymbol{x}')| - |f(\boldsymbol{y}')| \leq |f(\boldsymbol{x}') - f(\boldsymbol{y}')|$, so it follows that

$$|f(\boldsymbol{x}')| - |f(\boldsymbol{y}')| \leq \alpha\|\boldsymbol{x}' - \boldsymbol{y}'\|. \tag{4.5}$$

Rearranging, we find

$$|f(\boldsymbol{x}')| \leq |f(\boldsymbol{y}')| + \alpha\|\boldsymbol{x}' - \boldsymbol{y}'\|. \tag{4.6}$$

We conclude the proof by taking $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{y}$ and $\boldsymbol{y}' = \boldsymbol{x}$. By assumption, both $\boldsymbol{x}' = \boldsymbol{x} + \boldsymbol{y} \in B_r(\mathbf{0}, \|\cdot\|)$ and $\boldsymbol{y}' = \boldsymbol{x} \in B_r(\mathbf{0}, \|\cdot\|)$. Hence we find

$$|f(\boldsymbol{x} + \boldsymbol{y})| \leq |f(\boldsymbol{x})| + \alpha\|\boldsymbol{y}\|. \tag{4.7}$$

∎

We now define an inner radius δ by

$$\delta = \frac{1}{4\alpha} \min\{r, \gamma\}. \tag{4.8}$$

We see that $0 < \delta \leq r/4$ and $0 < \delta \leq \gamma/4$.

Lemma 4.1.2. *For any $\rho \in \mathcal{F}$ and $\boldsymbol{x} \in B_\delta(\mathbf{0}, \|\cdot\|)$, the function $f : [0, 1] \to \mathbb{R}$, $f(\lambda) = \rho(\lambda\boldsymbol{x})$ is continuous on $[0, 1]$, with $f(0) = 0$.*

Proof: We have that f is the composition of continuous functions (multiplication by a constant, followed by the Lipschitz function ρ), hence f is continuous. In addition, since $\rho(\mathbf{0}) = 0$, $f(0) = 0$. ∎

Lemma 4.1.3. *We let $\rho \in \mathcal{F}$ be an arbitrary function in $\mathcal{F}$. For all $\boldsymbol{x} \in B_\delta(\mathbf{0}, \|\cdot\|)$, let $\beta\|\boldsymbol{x}\| \leq \rho(\boldsymbol{x}) \leq \alpha\|\boldsymbol{x}\|$ and $\rho(\boldsymbol{x}) < \gamma/4$. Then for all $\boldsymbol{y} \in \mathbb{R}^d$ such that $\|\boldsymbol{y}\| \geq \delta$, we have $\rho(\boldsymbol{y}) \geq \beta\delta$.*

Proof: We let $\boldsymbol{x} \in B_\delta(\mathbf{0}, \|\cdot\|)$. If $\boldsymbol{x} = \mathbf{0}$, then both $\|\boldsymbol{x}\| = 0$ and $\rho(\boldsymbol{x}) = 0$ and we are done. We now suppose that $\boldsymbol{x} \neq \mathbf{0}$. We define $f(\lambda) = \rho(\lambda\boldsymbol{x})$, by assumption we have that $f'(\lambda) \geq \beta\|\boldsymbol{x}\|$ for all $\lambda \in (0, 1)$. By Lemma 4.1.2, f is continuous on $[0, 1]$, with $f(0) = \rho(\mathbf{0}) = 0$ and $f(1) = \rho(\boldsymbol{x})$. By the mean value theorem, there exists a $c \in (0, 1)$ such that $f(1) - f(0) = f'(c)(1 - 0)$, which implies that $f'(c) = f(1)$. We then find that

$$\begin{aligned}\rho(\boldsymbol{x}) &= f(1) \\ &= f'(c) \\ &\geq \beta\|\boldsymbol{x}\|.\end{aligned}$$

Since ρ is Lipschitz continuous on $B_\delta(\mathbf{0}, \|\cdot\|)$ with Lipschitz constant α, we also have that $\rho(\boldsymbol{x}) - \rho(\mathbf{0}) \leq \alpha(\|\boldsymbol{x}\| - \|\mathbf{0}\|)$, which implies (with $\rho(\mathbf{0}) = 0$) that $\rho(\boldsymbol{x}) \leq \alpha\|\boldsymbol{x}\|$. Hence we have that $\beta\|\boldsymbol{x}\| \leq \rho(\boldsymbol{x}) \leq \alpha\|\boldsymbol{x}\|$. We additionally notice that $\|\boldsymbol{x}\| < \delta$ and so

$$\begin{aligned}\rho(\boldsymbol{x}) &\leq \alpha\|\boldsymbol{x}\| \\ &< \alpha\delta \\ &\leq \alpha\frac{\gamma}{4\alpha} \\ &= \frac{\gamma}{4}.\end{aligned}$$

We let $\boldsymbol{y} \in \mathbb{R}^d$ be a point not in the open ball $B_\delta(\mathbf{0}, \|\cdot\|)$, so that $\|\boldsymbol{y}\| \geq \delta$. If $\|\boldsymbol{y}\| \geq r$, then $\rho(\boldsymbol{y}) \geq \gamma \geq \beta\gamma \geq \beta\delta$, and we are done. If $\delta \leq \|\boldsymbol{y}\| < r$, then we let $f(\lambda) = \rho(\lambda\boldsymbol{y})$, with f being differentiable on $(0,1)$ with $f'(\lambda) \geq \beta\|\boldsymbol{y}\|$ for all $\lambda \in (0,1)$. We also have (by Lemma 4.1.2) that f is continuous on $[0,1]$, $f(0) = 0$, and $f(1) = \rho(\boldsymbol{y})$. By the mean value theorem there exists a $c \in (0,1)$ such that $f(1) = f'(c)$. We then find that

$$\begin{aligned}\rho(\boldsymbol{y}) &= f(1)\\ &= f'(c)\\ &\geq \beta\|\boldsymbol{y}\|\\ &\geq \beta\delta.\end{aligned}$$

We conclude that if $\|\boldsymbol{y}\| \geq \delta$, then $\rho(\boldsymbol{y}) \geq \beta\delta$. ∎

Lemma 4.1.4. *For all points within $B_\delta(\mathbf{0}, \|\cdot\|)$, a triangle inequality with a α/β multiplicative constant holds, that is if $\|\boldsymbol{x}\| < \delta$ and $\|\boldsymbol{y}\| < \delta$, then*

$$\rho(\boldsymbol{x} + \boldsymbol{y}) \leq \frac{\alpha}{\beta}(\rho(\boldsymbol{x}) + \rho(\boldsymbol{y})). \tag{4.9}$$

Proof: We see that for all $\boldsymbol{x}, \boldsymbol{y} \in B_\delta(\mathbf{0}, \|\cdot\|)$, by Lemma 4.1.3,

$$\begin{aligned}\rho(\boldsymbol{x} + \boldsymbol{y}) &\leq \alpha\|\boldsymbol{x} + \boldsymbol{y}\|\\ &\leq \alpha(\|\boldsymbol{x}\| + \|\boldsymbol{y}\|)\\ &\leq \alpha\left(\frac{\rho(\boldsymbol{x})}{\beta} + \frac{\rho(\boldsymbol{y})}{\beta}\right)\\ &= \frac{\alpha}{\beta}\left(\rho(\boldsymbol{x}) + \rho(\boldsymbol{y})\right).\end{aligned}$$

∎

Lemma 4.1.5. *Suppose we have an iid query and sample points $X, X_1, X_2, \ldots, X_n$. We let $X_{(1,\|\cdot\|)}, X_{(2,\|\cdot\|)}, \ldots, X_{(n,\|\cdot\|)}$ be the points in increasing distance from X with respect to the norm $\|\cdot\|$ and $X_{(1,\rho)}, X_{(2,\rho)}, \ldots, X_{(n,\rho)}$ be the points in increasing distance from X with respect to the ρ distance, for any function $\rho \in \mathcal{F}$. We define the weight function $W_{ni}(X)$ to be $1/k$ if X_i is one of the k-nearest neighbours of X in the ρ distance (that is, if $X \in \{X_{(1,\rho)}, X_{(2,\rho)}, \ldots, X_{(k,\rho)}\}$, and to be zero otherwise. For any $a > 0$ and $\rho \in \mathcal{F}$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] \leq \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}\min\{a, \delta\}\right). \tag{4.10}$$

Proof: We show that if the condition on the right hand side of (4.10) does not hold, that is, if $\left\|X_{(k,\|\cdot\|)} - X\right\| \leq \frac{\beta}{2\alpha}\min\{a, \delta\}$, then $\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}} = 0$. We assume that $\left\|X_{(k,\|\cdot\|)} - X\right\| \leq \frac{\beta}{2\alpha}\min\{a, \delta\}$, since this is less than δ by assumption we can use our above results. We first define $a' = \min\{a, \delta\}$. We let $i \in \{1, 2, \ldots, k\}$ and observe that

$$\begin{aligned}
\rho(X - X_{(i,\rho)}) &\leq \alpha\left\|X - X_{(i,\rho)}\right\| \\
&\leq \alpha\left\|X - X_{(k,\rho)}\right\| \\
&\leq \alpha\frac{\beta}{2\alpha}a' \\
&= \frac{\beta}{2}a'.
\end{aligned}$$

This means there are at least k points such that the ρ-distance from X to the point is at most $\frac{\beta}{2}a'$. Suppose we have a point X_j such that $\|X - X_j\| > a'$. There are two possible cases: either $\|X - X_j\| \geq \delta$ or $a' < \|X - X_j\| < \delta$ (only the first case is possible if $a' = \delta$). If $\|X - X_j\| \geq \delta$, by Lemma 4.1.3 we have that $\rho(X - X_j) \geq \beta\delta \geq \beta a'$. If $a' < \|X - X_j\| < \delta$, then $\rho(X - X_j) \geq \beta\|X - X_j\| > \beta a'$. In both cases $\rho(X - X_j) \geq \beta a' > \frac{\beta}{2}a' \geq \rho(X - X_{(i,\rho)})$, for each $1 \leq i \leq k$. From this we see that there are at least k points closer to X than X_j in the ρ distance.

From this we see that if $\left\|X_{(k,\|\cdot\|)} - X\right\| \leq \frac{\beta}{2\alpha}\min\{a, \delta\}$ holds, then for all points X_i such that $\|X_i - X\| > a$, X_i is not a k-nearest neighbour of X in the ρ distance,

and since $W_{ni}(X)$ is only nonzero for the k-nearest neighbours of X in the ρ distance, it follows that

$$\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}} = 0.$$

We notice that $\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}} \leq 1$ always and is nonnegative (since $\sum_{i=1}^{n} W_{ni}(X) = 1$ and the indicator function is either zero or one). We then condition on $\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}a'$ to find that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}} \middle| \left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}a'\right] \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}a'\right) + \\
&\quad \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}} \middle| \left\|X_{(k,\|\cdot\|)} - X\right\| \leq \frac{\beta}{2\alpha}a'\right] \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| \leq \frac{\beta}{2\alpha}a'\right) \\
&\leq (1)\left(\mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}a'\right)\right) + 0 \\
&= \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}\min\{a, \delta\}\right).
\end{aligned}$$

∎

Lemma 4.1.6. *For any $a > 0$ and $\epsilon > 0$, there exists an $N \geq 1$ such that for all $n \geq N$, for all $\rho \in \mathcal{F}$, with $W_{ni}(X)$ as defined in Lemma 4.1.5,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] < \epsilon. \tag{4.11}$$

Proof: By Lemma 4.1.5, we have that for all $\rho \in \mathcal{F}$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] \leq \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}\min\{a, \delta\}\right). \tag{4.12}$$

We define $a' = \frac{\beta}{2\alpha}\min\{a, \delta\}$ (we notice that δ is a fixed constant for the family $\mathcal{F}$, defined by (4.8)). By Lemma 2.3.4, $\mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > a'\right) \to 0$ as $n \to \infty$, so

there exists an $N \geq 1$ such that for all $n \geq N$,

$$\mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha}\min\{a, \delta\}\right) < \epsilon. \tag{4.13}$$

Hence it follows from (4.12) and (4.13) that for all $\rho \in \mathcal{F}$ and $n \geq N$,

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] < \epsilon. \tag{4.14}$$

∎

Lemma 4.1.7. *There is a uniform upper bound $c > 0$ such that for any $\rho \in \mathcal{F}$, there are at most c subsets $S_1, S_2, \ldots, S_n$ covering $B_\delta(\mathbf{0}, \|\cdot\|)$ such that for each S_i (where $1 \leq i \leq c$), if $\boldsymbol{x}, \boldsymbol{y} \in S_i$ with $\boldsymbol{x} \neq \mathbf{0}$ and $\rho(\boldsymbol{x}) \leq \rho(\boldsymbol{y})$, then $\rho(\boldsymbol{y} - \boldsymbol{x}) < \rho(\boldsymbol{y})$.*

Proof: Since $B_\delta(\mathbf{0}, \|\cdot\|)$ is a bounded subset of $\mathbb{R}^d$ there exists a covering with c open balls of radius $\beta^2/(4\alpha^2)$ each in the $\|\cdot\|$ norm (this follows from the fact that any bounded subset of the normed space $(\mathbb{R}^d, \|\cdot\|)$ is precompact or totally bounded). We then find that for any $\boldsymbol{v} \in B_\delta(\mathbf{0}, \|\cdot\|)$ (using Lemma 4.1.3),

$$\begin{aligned} 0 &\leq \left\|\frac{\boldsymbol{v}}{\rho(\boldsymbol{v})}\right\| \\ &= \frac{1}{\rho(\boldsymbol{v})}\|\boldsymbol{v}\| \\ &\leq \frac{1}{\beta\|\boldsymbol{v}\|}\|\boldsymbol{v}\| \\ &\leq \frac{1}{\beta}. \end{aligned}$$

This means that for any vector $\boldsymbol{v} \in B_\delta(\mathbf{0}, \|\cdot\|) \setminus \{\mathbf{0}\}$, $\frac{\boldsymbol{v}}{\rho(\boldsymbol{v})} \in B_{2/\beta}(\mathbf{0}, \|\cdot\|)$ (since $0 \leq \left\|\frac{\boldsymbol{v}}{\rho(\boldsymbol{v})}\right\| < 2/\beta$). We then let $T_1, T_2, \ldots, T_c$ be the open balls that cover $B_{2/\beta}(\mathbf{0}, \|\cdot\|)$ with radius $\beta^2/(4\alpha^2)$ each and $S_1, S_2, \ldots, S_c$ be the subsets of $B_\delta(\mathbf{0}, \|\cdot\|)$ such that $\boldsymbol{v} \in S_i$ if and only if either $\boldsymbol{v} = \mathbf{0}$ or $\boldsymbol{v} \neq \mathbf{0}$ and $\frac{\boldsymbol{v}}{\rho(\boldsymbol{v})} \in T_i$. We see that every vector $\boldsymbol{v} \in B_\delta(\mathbf{0}, \|\cdot\|)$ is in at least one S_i since the zero vector is in every S_i and every

nonzero such vector has $\frac{\boldsymbol{v}}{\rho(\boldsymbol{v})} \in T_i$ for some $1 \leq i \leq c$ so $\boldsymbol{v} \in S_i$. For each T_i we let $\boldsymbol{x}_i \in T_i$ be an element of T_i.

Suppose that $\boldsymbol{x}, \boldsymbol{y} \in S_i$ with $\rho(\boldsymbol{x}) \leq \rho(\boldsymbol{y})$ and $\rho(\boldsymbol{x}) > 0$. Since $\boldsymbol{x}, \boldsymbol{y} \in S_i$ we have that $\boldsymbol{x}/\|\boldsymbol{x}\|, \boldsymbol{y}/\|\boldsymbol{y}\| \in T_i$, and since $\boldsymbol{x}_i \in T_i$ we have that $\left\|\frac{\boldsymbol{y}}{\rho(\boldsymbol{y})} - \boldsymbol{x}_i\right\| < \frac{\beta^2}{4\alpha^2}$ and that $\left\|\boldsymbol{x}_i - \frac{\boldsymbol{x}}{\rho(\boldsymbol{x})}\right\| < \frac{\beta^2}{4\alpha^2}$. We then find that

$$\begin{aligned}
\left\|\frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} - \boldsymbol{x}\right\| &= \rho(\boldsymbol{x})\left\|\frac{\boldsymbol{y}}{\rho(\boldsymbol{y})} - \frac{\boldsymbol{x}}{\rho(\boldsymbol{x})}\right\| \\
&= \rho(\boldsymbol{x})\left\|\frac{\boldsymbol{y}}{\rho(\boldsymbol{y})} - \boldsymbol{x}_i + \boldsymbol{x}_i - \frac{\boldsymbol{x}}{\rho(\boldsymbol{x})}\right\| \\
&\leq \rho(\boldsymbol{x})\left(\left\|\frac{\boldsymbol{y}}{\rho(\boldsymbol{y})} - \boldsymbol{x}_i\right\| + \left\|\boldsymbol{x}_i - \frac{\boldsymbol{x}}{\rho(\boldsymbol{x})}\right\|\right) \\
&< \rho(\boldsymbol{x})\left(\frac{\beta^2}{4\alpha^2} + \frac{\beta^2}{4\alpha^2}\right) \\
&= \frac{\beta^2}{2\alpha^2}\rho(\boldsymbol{x}).
\end{aligned}$$

We notice that $\boldsymbol{y} - \boldsymbol{x} = \boldsymbol{y} - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} + \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} - \boldsymbol{x}$. We see that $\|\boldsymbol{x}\| < \delta$, $\|\boldsymbol{y}\| < \delta$, and that

$$\begin{aligned}
\left\|\frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y}\right\| &= \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\|\boldsymbol{y}\| \\
&\leq \|\boldsymbol{y}\| \\
&< \delta.
\end{aligned}$$

From this we see that the norm of the first part is less than r,

$$\begin{aligned}
\left\|\boldsymbol{y} - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y}\right\| &\leq \|\boldsymbol{y}\| + \left\|\frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\right\| \\
&< \delta + \delta \\
&\leq r/2.
\end{aligned}$$

The second part has a norm less than r,

$$\left\|\frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} - \boldsymbol{x}\right\| \leq \|\boldsymbol{x}\| + \left\|\frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y}\right\|$$

$$< \delta + \delta$$
$$\leq r/2.$$

The norm of sum of both parts is less than r,

$$\left\| \boldsymbol{y} - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} + \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} - \boldsymbol{x} \right\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\| + 2\left\| \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} \right\|$$
$$< 4\delta$$
$$\leq r.$$

We are now able to apply Lemma 4.1.1 and our above observations to find that (along with the fact that since $\alpha \geq 1$ and $\beta > 0$, $\frac{\beta^2}{2\alpha} - \beta^2 < 0$)

$$\begin{aligned}
\rho(\boldsymbol{y} - \boldsymbol{x}) &= \rho\left(\boldsymbol{y} - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} + \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} - \boldsymbol{x} \right) \\
&\leq \rho\left(\boldsymbol{y} - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} \right) + \alpha \left\| \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} - \boldsymbol{x} \right\| \\
&< \rho\left(\boldsymbol{y} - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} \right) + \alpha \frac{\beta^2}{2\alpha^2}\rho(\boldsymbol{x}) \\
&= \rho\left(\boldsymbol{y} - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\boldsymbol{y} \right) + \frac{\beta^2}{2\alpha}\rho(\boldsymbol{x}) \\
&\leq \rho(\boldsymbol{y}) - \frac{\rho(\boldsymbol{x})}{\rho(\boldsymbol{y})}\beta^2\rho(\boldsymbol{y}) + \frac{\beta^2}{2\alpha}\rho(\boldsymbol{x}) \\
&= \rho(\boldsymbol{y}) + \left(\frac{\beta^2}{2\alpha} - \beta^2 \right)\rho(\boldsymbol{x}) \\
&\leq \rho(\boldsymbol{y}).
\end{aligned}$$

This means that for all nonzero $\boldsymbol{x}, \boldsymbol{y} \in S_i$ with $\rho(\boldsymbol{x}) \leq \rho(\boldsymbol{y})$, $\rho(\boldsymbol{y} - \boldsymbol{x}) < \rho(\boldsymbol{y})$. ∎

Lemma 4.1.8. *Let X be a query and $X_1, X_2, \ldots, X_n$ be the sample points, all of which are iid. Let $(\rho_n)_{n=1}^{\infty}$ be a sequence of functions in $\mathcal{F}$ (that is possibly random, but is independent of the labelled sample and the query), and W_{ni} be the corresponding weights. There exists a constant $c > 0$ (with the c defined in Lemma 4.1.7) and a*

sequence $\epsilon_n \to 0$ *as* $n \to \infty$ *such that for any nonnegative measurable function* f *bounded above by one,*

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}\left[f(X)\right] + \epsilon_n. \tag{4.15}$$

Proof: We notice that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)(\mathbb{1}_{\{\|X_i - X\| < \delta\}} + \mathbb{1}_{\{\|X_i - X\| \geq \delta\}})\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\mathbb{1}_{\{\|X_i - X\| < \delta\}}\right] + \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\mathbb{1}_{\{\|X_i - X\| \geq \delta\}}\right].
\end{aligned}$$

For the first term, we have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\mathbb{1}_{\{\|X_i - X\| < \delta'\}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{k}\mathbb{1}_{\{X_i \text{ is a } \rho_n \text{ } k\text{-NN of } X \text{ among } X_1,\ldots,X_i,\ldots,X_n\}} f(X_i)\mathbb{1}_{\{\|X_i - X\| < \delta\}}\right] \\
&= \frac{1}{k}\mathbb{E}\left[\sum_{i=1}^{n} f(X_i)\mathbb{1}_{\{X_i \text{ is a } \rho_n \text{ } k\text{-NN of } X \text{ among } X_1,\ldots,X_i,\ldots,X_n \text{ and } \|X_i - X\| < \delta\}}\right] \\
&= \frac{1}{k}\mathbb{E}\left[\sum_{i=1}^{n} f(X)\mathbb{1}_{\{X \text{ is a } \rho_n \text{ } k\text{-NN of } X_i \text{ among } X_1,\ldots,X,\ldots,X_n \text{ and } \|X - X_i\| < \delta\}}\right] \\
&= \frac{1}{k}\mathbb{E}\left[f(X)\sum_{i=1}^{n} \mathbb{1}_{\{X \text{ is a } \rho_n \text{ } k\text{-NN of } X_i \text{ among } X_1,\ldots,X,\ldots,X_n \text{ and } \|X - X_i\| < \delta\}}\right].
\end{aligned}$$

We define subsets $S_1, S_2, \ldots, S_c$ as in Lemma 4.1.7. In each of the subsets $S_1, S_2, \ldots, S_c$, we mark the k points closest to X in the ρ_n distance (we recall that these subsets cover $B_a(\mathbf{0}, \|\cdot\|)$). Suppose the point X_i in the subset S_q is not marked. Then there are at least k points $X_{j_1}, X_{j_2}, \ldots, X_{j_k}$ in S_q such that $\rho_n(X_{j_l} - X) \leq \rho_n(X_i - X)$ that are marked. For these points, if $X_{j_l} \neq X$, then $\rho_n(X_i - X_{j_l}) < \rho_n(X_i - X)$ by Lemma 4.1.7, and if $X_{j_l} = X$, then $\rho_n(X_i - X_{j_l}) = \rho_n(X_i - X)$ and $U_i < U_{j_l}$ must hold

(the U_i being the tiebreaking variables), in both cases X_{j_l} is selected as being closer to X_i in the ρ_n distance. Hence there are at least k points closer in the ρ_n distance to X_i than X. This means that if X_i is not marked and $\|X_i - X\| < a$, then X is not a k-nearest neighbour of X_i among $X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n$. Furthermore, the number of points that are marked is at most ck. It follows that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i) \mathbb{1}_{\{\|X_i - X\| < \delta\}}\right] &\leq \frac{1}{k}\mathbb{E}\left[f(X)\sum_{i=1}^{n}\mathbb{1}_{\{X_i \text{ is marked}\}}\right] \\
&\leq \frac{1}{k}\mathbb{E}\left[f(X)ck\right] \\
&\leq c\mathbb{E}\left[f(X)\right].
\end{aligned}$$

For the second term, we see that since f is nonnegative and bounded above by 1 and by Lemma 4.1.6,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i) \mathbb{1}_{\{\|X_i - X\| \geq \delta\}}\right] &\leq \mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| \geq \delta\}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{k} \frac{1}{k} \mathbb{1}_{\left\{\left\|X_{(i,\rho_n)} - X\right\| \geq \delta\right\}}\right] \\
&\leq \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\left[\mathbb{1}_{\left\{\left\|X_{(k,\rho_n)} - X\right\| \geq \delta\right\}}\right] \\
&= \mathbb{E}\left[\mathbb{1}_{\left\{\left\|X_{(k,\rho_n)} - X\right\| \geq \delta\right\}}\right] \\
&= \epsilon_n.
\end{aligned}$$

Combining these results, we find that

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}\left[f(X)\right] + \epsilon_n.$$

∎

We are now able to prove our main theorem, that k-NN with a sequence of functions in $\mathcal{F}$ chosen independently of the sample and the query is universally consistent. We restate the conditions on $\mathcal{F}$ to make the statement of the theorem self-contained.

Theorem 4.1.9. *Let $\mathcal{F}$ be a family of measurable functions from $\mathbb{R}^d$ to $\mathbb{R}$ such that there exist constants $\alpha \geq 0$, $\beta > 0$, $\gamma > 0$, and $r > 0$ so that for each function $\rho \in \mathcal{F}$,*

1. *The function ρ is α-Lipschitz on $B_r(\mathbf{0}, \|\cdot\|)$.*
2. *For any $\boldsymbol{x} \in B_r(\mathbf{0}, \|\cdot\|) \setminus \{\mathbf{0}\}$, the derivative of $f(\lambda) = \rho(\lambda \boldsymbol{x})$ is bounded below by $\beta\|\boldsymbol{x}\|$ for all $\lambda \in (0,1)$ (that is, $f'(\lambda) \geq \lambda\|\boldsymbol{x}\|$).*
3. *Outside $B_r(\mathbf{0}, \|\cdot\|)$, the function is bounded below by γ, so if $\|\boldsymbol{x}\| \geq r$, $\rho(\boldsymbol{x}) \geq \gamma$.*
4. *The function ρ is symmetric, so $\rho(\boldsymbol{x}) = \rho(-\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$.*
5. *The function ρ takes value zero at the zero vector, $\rho(\mathbf{0}) = 0$.*

Let $(\rho_n)_{n=1}^{\infty}$ be any sequence of random functions in $\mathcal{F}$, independent of the sample and query (with n being the sample size). We have that k-NN with the sequence of random functions $(\rho_n)_{n=1}^{\infty}$ is universally consistent (where given points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we take $\rho_n(\boldsymbol{x}, \boldsymbol{y})$ to be the distance between these points for k-NN). Furthermore, if we have a family of functions $\mathcal{G}$ such that every function in $\mathcal{G}$ is the composition of a function in $\mathcal{F}$ with a strictly increasing function (that is, for any $g \in \mathcal{G}$, $g = h \circ f$, with $f \in \mathcal{F}$ and h being strictly increasing), then k-NN with an sequence of random functions in $\mathcal{G}$ independent of the sample and the query is universally consistent.

Proof: We first notice that if we increase α to any larger value and decrease β to any smaller value greater than 0 in the definition of $\mathcal{F}$, any function in $\mathcal{F}$ remains inside. Hence we may assume without loss of generality that $\alpha \geq 1$ and $\beta \in (0, 1]$.

Let $(\rho_n)_{n=1}^{\infty}$ be a sequence of functions in $\mathcal{F}$. We see by Lemma 4.1.8 that for any nonnegative function f bounded above by one,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] &= \frac{1}{k}\mathbb{E}\left[\sum_{i=1}^{k} f(X_{(i,\rho_n)})\right] \\ &\leq \frac{1}{k}\left(kc\mathbb{E}\left[f(X)\right] + k\epsilon_n\right)\end{aligned}$$

$$= c\mathbb{E}\left[f(X)\right] + \epsilon_n$$

and so the first condition of Stone's theorem is satisfied. By Lemma 4.1.6, the second condition of Stone's theorem is satisfied. Since $k \to \infty$ as $n \to \infty$, $\frac{1}{k} \to 0$ as $n \to \infty$ and so the third condition of Stone's theorem holds. Hence, k-NN with any sequence of random functions in $\mathcal{F}$ (such that the sequence is independent of the sample and the query) is universally consistent, and so k-NN with a sequence of random functions in $\mathcal{F}$ (independent of the sample and query) is universally consistent. The choice of random function from the family can depend on the sample size n.

By Lemma 2.3.10, we have that since the transformations we apply after the distance are strictly increasing, the result of k-NN remains the same, so k-NN is universally consistent with a sequence of random functions (independent of the sample and the query) in $\mathcal{G}$ as well as $\mathcal{F}$. ∎

Remark 4.1.10. *We notice that functions $\rho \in \mathcal{F}$ may take on the special value ∞ outside the ball $B_r(\mathbf{0})$, which is larger than any finite value. The universal consistency proof holds in the same manner in this case.*

4.2 Families of Lipschitz Distances

In this section, we build families of Lipschitz distances that satisfy the conditions of Theorem 4.1.9.

Lemma 4.2.1. *Let $f : \mathbb{R}^+ \to \mathbb{R}$ be a function and $\alpha > 0$, $r > 0$ be constants such that $f(0) = 0$, f is continuous on $[0, r]$ and differentiable on $(0, r)$, with $f'(x) \leq \alpha$ for all $x \in (0, r)$. We then have that f is α-Lipschitz on $[0, r]$, and the function $g(x) = f(|x|)$ is α-Lipschitz on $[-r, r]$.*

Proof: This result follows from the intermediate value theorem. ∎

Lemma 4.2.2. *Let $\mathcal{F}$ be a family of functions from $\mathbb{R}^d$ to $\mathbb{R}$ and $\alpha > 0$, $\beta > 0$, $\gamma > 0$, and $r > 0$ be constants such that each $\rho \in \mathcal{F}$, we have*

$$\rho((x_1, x_2, \ldots, x_d)) = \sum_{i=1}^{d} f_i(|x_i|) \tag{4.16}$$

where each function $f_i : \mathbb{R}^+ \to \mathbb{R}^+$ is such that $f_i(x) = 0$, $\beta < f_i'(x) < \alpha$ for all $x \in (0, r)$, and $f_i(x) \geq \gamma$ for all $x \geq r$. Then k-NN with q sequence of random functions in $\mathcal{F}$ (independent of the sample and the query) is universally consistent.

Proof: We show that the family of distances $\mathcal{F}$ satisfies the conditions of Theorem 4.1.9. For the first condition, we let $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)$ be two points in $B_r(\mathbf{0}, \|\cdot\|)$, by expanding $|\rho(\boldsymbol{x}) - \rho(\boldsymbol{y})|$ and applying Lemma 4.2.1 we find

$$\begin{aligned}
|\rho(\boldsymbol{x}) - \rho(\boldsymbol{y})| &= \left| \sum_{i=1}^{d} (f_i(|x_i|) - f_i(|y_i|)) \right| \\
&\leq \sum_{i=1}^{d} |(f_i(|x_i|) - f_i(|y_i|))| \\
&\leq \sum_{i=1}^{d} \alpha \, |x_i - y_i| \\
&= \alpha \|\boldsymbol{x} - \boldsymbol{y}\|_1.
\end{aligned}$$

Hence we have that every function in $\mathcal{F}$ is α-Lipschitz in the $\|\boldsymbol{x} - \boldsymbol{y}\|_1$ on $B_r(\mathbf{0}, \|\cdot\|)$. For the second condition, we see that for all $\lambda \in (0, 1)$ and $\boldsymbol{x} \neq \mathbf{0}$ with $\|\boldsymbol{x}\| \leq r$,

$$\begin{aligned}
\frac{\partial}{\partial \lambda} \rho(\lambda \boldsymbol{x}) &= \frac{\partial}{\partial \lambda} \sum_{i=1}^{d} f_i(|\lambda x_i|) \\
&= \sum_{i=1}^{d} |x_i| \, f_i'(|\lambda x_i|) \\
&\geq \beta \sum_{i=1}^{d} |x_i|
\end{aligned}$$

$$= \beta \|\boldsymbol{x}\|_1.$$

For the third condition, we notice that for all $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\|_\infty \geq r$, the minimum $\min\{|x_1|, |x_2|, \ldots, |x_d|\} \geq r$, which implies $f_i(|\lambda x_i|) \geq \gamma$ for some $i \in \{1, 2, \ldots, d\}$, and hence $\rho(\boldsymbol{x}) \geq \gamma$. The fourth condition follows directly from the fact we take the absolute value of each of the x_i, and the fifth condition holds since $\rho(\mathbf{0}) = \sum_{i=1}^d f_i(0) = 0$. ∎

Corollary 4.2.3. *For each of the following functions, k-NN with any of the distances* $\rho(\boldsymbol{x}) = \sum_{i=1}^d f(|x_i|)$ *is universally consistent:*

1. *The exponential function* $f_1(x) = e^x$.
2. *The function* $f_2(x) = \begin{cases} \sin(x) & \text{if } x \leq 1 \\ x & \text{if } x \geq 1 \end{cases}$.
3. *The function* $f_3(x) = \begin{cases} \tan(x) & \text{if } x < \pi/2 \\ \infty & \text{if } x \geq \pi/2 \end{cases}$.
4. *The arctangent function* $f_4(x) = \arctan(x)$.
5. *The hyperbolic sine function* $f_5(x) = \sinh(x)$.
6. *The hyperbolic tangent function* $f_6(x) = \tanh(x)$.

Proof: We see that Lemma 4.2.2 hold for the functions f_2, f_3, f_4, f_5, f_6 directly, and $e^x - 1$ satisfies the conditions of the lemma and is a strictly increasing transformation of f_1. ∎

We now show that Theorem 3.2.5 follows from Theorem 4.1.9.

Theorem 4.2.4. *Let $\mathcal{F}$ be a family of norms on $\mathbb{R}^d$ such that there exists a norm $\|\cdot\|$ on $\mathbb{R}^d$ and a constant $C \geq 1$ such that for all $\rho \in \mathcal{F}$ and $\boldsymbol{x} \in \mathbb{R}^d$, $\frac{1}{C}\|\boldsymbol{x}\| \leq \rho(\boldsymbol{x}) \leq C\|\boldsymbol{x}\|$. We then have that k-NN with any sequence of random norms (independent of the sample and the query) in $\mathcal{F}$ is universally consistent.*

Proof: We let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ be two points in $\mathbb{R}^d$. We see that

$$\begin{aligned} |\rho(\boldsymbol{x}) - \rho(\boldsymbol{y})| &\leq \rho(\boldsymbol{x} - \boldsymbol{y}) \\ &\leq C\|\boldsymbol{x} - \boldsymbol{y}\|. \end{aligned}$$

Hence we have that ρ is Lipschitz with constant C on $(\mathbb{R}^d, \|\cdot\|)$. Furthermore, we see that if we let $f(\lambda) = \rho(\lambda \boldsymbol{x})$ for some $\boldsymbol{x} \neq 0$, for all $\lambda \in (0, 1)$,

$$\begin{aligned} f'(\lambda) &= \frac{\partial}{\partial \lambda}\rho(\lambda \boldsymbol{x}) \\ &\leq \frac{\partial}{\partial \lambda}\frac{1}{C}\|\lambda \boldsymbol{x}\| \\ &= \frac{\partial}{\partial \lambda}\frac{|\lambda|}{C}\|\boldsymbol{x}\| \\ &= \frac{1}{C}\|\boldsymbol{x}\|. \end{aligned}$$

Additionally, we have that for all $\boldsymbol{x}$ such that $\|\boldsymbol{x}\| \geq 1$, $\rho(\boldsymbol{x}) \geq \frac{1}{C}\|\boldsymbol{x}\| \geq \frac{1}{C}$. We also see that $\rho(\boldsymbol{x}) = \rho(-\boldsymbol{x})$ and $\rho(\mathbf{0}) = 0$ since ρ is a norm. Hence we find that $\mathcal{F}$ satisfies the conditions of Theorem 4.1.9 with $\alpha = C$, $\beta = \frac{1}{C}$, $\gamma = \frac{1}{C}$, and $r = 1$. ∎

Unfortunately, this result does not extend to quasinorms on $\mathbb{R}^d$, as they are not necessarily Lipschitz (or even locally Lipschitz near zero). In particular, the ℓ^p quasinorms with $p \in (0, 1)$ are not Lipschitz for any open ball around zero, with their partial derivatives in the i^{th} coordinate being unbounded near the i^{th} axis. To prove universal consistency for k-NN with quasinorms (which remains an open question), one must employ a different approach.

We can also consider families of polynomials.

Theorem 4.2.5. *Let $\alpha, \beta > 0$ be fixed constants and $p \geq 1$ be an integer. Let $\mathcal{F}$ be the family of polynomials of the form*

$$\mathcal{F} = \left\{ a_1 x + a_2 x^2 + \cdots + a_p x^p \;\middle|\; a_1 \geq \beta;\ \forall m \geq 1, 0 \leq a_m \leq \alpha \right\}. \tag{4.17}$$

That is, for any polynomial in $\mathcal{F}$ the first coefficient must be at least β and all coefficients must be nonnegative and at most α. Suppose we let $\mathcal{G}$ be the family of functions defined by applying f to the modulus of the difference of each coordinate. Then k-NN with any sequence of random functions in $\mathcal{G}$ (independent of the sample and the query) is universally consistent.

Proof: We verify the conditions of Lemma 4.2.2 for any function $f \in \mathcal{F}$. First, we find an upper bound for the derivative of f on the interval $[0, 1]$,

$$\begin{aligned} f'(x) &= \sum_{m=1}^{p} a_m m x^{m-1} \\ &\leq \alpha \sum_{m=1}^{p} m \\ &= \alpha \frac{m(m+1)}{2}. \end{aligned}$$

We also find that f' is bounded below by β on $(0, \infty)$, since $f'(x) = \sum_{m=1}^{p} a_m m x^{m-1}$ with the first term being β and all the other terms being nonnegative. This implies that f is monotone increasing on $(0, \infty)$, so $f(x) \geq f(1)$ for all $x \geq 1$. Hence we have that f satisfies the conditions of Lemma 4.2.2 and so k-NN with a sequence from the corresponding family of distances $\mathcal{G}$ is universally consistent. ∎

As with norms, we can take linear combinations of these functions and add them to our family.

Lemma 4.2.6. *Let $\mathcal{F}$ be a family of distances satisfying the conditions of Theorem 4.1.9. If we let $\mathcal{G}$ be the family of all functions $\rho : \mathbb{R}^d \to \mathbb{R}$ of the form (with $p \geq 1$*

being a fixed constant)

$$\rho((x_1, x_2, \ldots, x_d)) = \sum_{i=1}^{p} \frac{a_i}{A} \rho_i(\boldsymbol{x}) \tag{4.18}$$

where $0 \leq a_i \leq 1$*, at least one* a_i *is strictly positive* ($a_i > 0$)*,* $A = \sum_{i=1}^{p} a_i$*, and* $\rho_i \in \mathcal{F}$*. Then* k*-NN any sequence of functions in* $\mathcal{G}$ *(independent of the sample and the query) is universally consistent (in* $\mathcal{G}$*, we can also include strictly increasing functions of such linear combinations).*

Proof: For any function $\rho \in \mathcal{G}$ of the above form, we see that ρ is αp-Lipschitz, since

$$\begin{aligned}
|\rho(\boldsymbol{x}) - \rho(\boldsymbol{y})| &= \left| \sum_{i=1}^{p} \frac{a_i}{A} (\rho_i(\boldsymbol{x}) - \rho_i(\boldsymbol{y})) \right| \\
&\leq \sum_{i=1}^{p} \frac{a_i}{A} |\rho_i(\boldsymbol{x}) - \rho_i(\boldsymbol{y})| \\
&\leq \sum_{i=1}^{p} |\rho_i(\boldsymbol{x}) - \rho_i(\boldsymbol{y})| \\
&= \alpha p \|\boldsymbol{x} - \boldsymbol{y}\|.
\end{aligned}$$

Furthermore, we see that the derivative is bounded below by $\beta p \|\boldsymbol{x}\|$,

$$\begin{aligned}
\frac{\partial}{\partial \lambda} \rho(\lambda \boldsymbol{x}) &= \frac{\partial}{\partial \lambda} \sum_{i=1}^{p} \frac{a_i}{A} (\rho_i(\lambda \boldsymbol{x})) \\
&\geq \sum_{i=1}^{p} \frac{a_i}{A} \beta \|\boldsymbol{x}\| \\
&= \beta p \|\boldsymbol{x}\|.
\end{aligned}$$

Since each ρ_i is bounded below by γ outside $B_r(\mathbf{0})$ for some fixed $r > 0$, we see that for all $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\boldsymbol{x}\| \geq r$,

$$\rho(\boldsymbol{x}) = \sum_{i=1}^{p} \frac{a_i}{A} \rho_i(\boldsymbol{x})$$

$$\geq \sum_{i=1}^{p} \frac{a_i}{A} \gamma$$

$$= \gamma.$$

For the fourth and fifth conditions, we see that since each ρ_i is symmetric and takes value zero at the zero vector, the same holds for ρ.

By Lemma 2.3.10, we can include strictly increasing transformations of such functions, and the same result will be generated, so universal consistency is preserved. ∎

Chapter 5

Adaptive k-NN

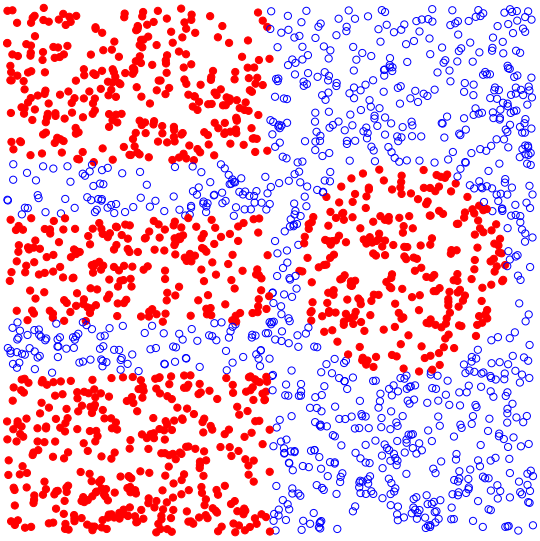

Figure 5.1: We see we would like to use a different norm for k-NN on the left side and the right side, and possibly within the right side, for this dataset.

In this chapter, we investigate under what conditions we can select the distance for k-NN based on the query and the sample and retain the universal consistency proof in Stone's theorem. We first look at a limitation of Stone's theorem, which prevents us from using the sample labels for a classifier we would like to prove is

universal consistent using Stone's theorem. We then define a modified k-NN in which we can select the distance based on the query and on the sample points (but not the labels). We illustrate why such an adaptive procedure is useful in Figure 5.1.

5.1 Limitations of Stone's Theorem in Adaptive k-NN

In Stone's theorem, we assumed that the weights $W_{ni}(X)$ are a functions of the query X, the sample points $X_1, X_2, \ldots, X_n$, and an independent random variable V only, and not on the sample labels $Y_1, Y_2, \ldots, Y_n$. In the following example we show that this assumption is necessary.

Example 5.1.1. Suppose we have a joint distribution for $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ with X having a multivariate normal distribution and Y being an independent Bernoulli random variable with a probability of $1/3$ of being zero and $2/3$ of being one. We have an iid labelled sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ and a query (X, Y) from this distribution. It is clear that the Bayes error for this distribution is $1/3$, and is attained by selecting label one always.

Suppose we let $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. We define weights $W_{ni}(X)$ that are $1/k$ if X_i is both a $5k$ nearest neighbour of X in the Euclidean norm and X_i is one of the k nearest points to X among those $5k$ points with label zero (if there are not enough such points, then we take points among the nearest $5k$ with label one until we have sufficiently many points), and zero otherwise. We see that the weights $W_{ni}(X)$ depend on the labels $Y_1, Y_2, \ldots, Y_n$ and that they can only be nonzero for points that among the $5k$ nearest neighbours of X in the Euclidean norm. By the strong law of large numbers the fraction of the nearest $5k$ points with label zero approaches $1/3$ as $n \to \infty$. Hence with probability approaching one there will be at least k points among the nearest $5k$ with label zero, thus the weights will be nonzero only for points

with label zero, and so the point will be classified as zero by the weight based classifier (of the form in equation (2.3)). It follows that the expected misclassification error is $2/3$, which is much higher than the Bayes error (indeed, it is worse than randomly guessing).

However, we see that $W_{ni}(X)$ satisfies the three conditions of Stone's theorem (the only assumption violated is the initial one that $W_{ni}(X)$ does not depend on the sample labels):

1. For the first condition, from Lemma 2.3.8 we find that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{k} \mathbb{1}_{\{X_i \text{ is a } \|\cdot\| \text{ } 5k\text{-NN of } X \text{ among } X_1,\ldots,X_{i-1},X_i,X_{i+1},\ldots,X_n\}} f(X_i)\right] \\
&= \frac{1}{k}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{X \text{ is a } \|\cdot\| \text{ } 5k\text{-NN of } X_i \text{ among } X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}} f(X)\right] \\
&\leq 5c\mathbb{E}[f(X)].
\end{aligned}$$

2. For the second condition, we see that $\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}$ is bounded above by one, is nonnegative, and is nonzero if and only if the $5k^{\text{th}}$ nearest point to X has a distance of at most a. By Lemma 2.3.4, the probability of this goes to zero as n goes to infinity (since $5k/n \to 0$ as $n \to \infty$), and hence the expected value $\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X)\mathbb{1}_{\{\|X_i - X\| > a\}}\right] \to 0$ as $n \to \infty$.

3. The third condition follows from the fact that $k \to \infty$ as $n \to \infty$, so that $\max_{1 \leq i \leq n} W_{ni}(X) = 1/k \to 0$ as $n \to \infty$.

This example satisfies the three conditions of Stone's theorem, but is not universally consistent, the only assumption violated is that the weights $W_{ni}(X)$ depends on the sample labels $Y_1, Y_2, \ldots, Y_n$. Hence this requirement in Stone's theorem is essential and cannot be removed. If a learning rule depends on the sample labels

$Y_1, Y_2, \ldots, Y_n$ for the weights $W_{ni}(X)$, we cannot use Stone's theorem (at least without modifications for that specific learning rule) to prove it is universally consistent.

5.2 Consistency of k-NN with an Adaptively Chosen Sequence of Distances

In this section we investigate the conditions under which we can select the distance for k-NN, depending on the query X and the sample points $X_1, X_2, \ldots, X_n$.

Theorem 5.2.1. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$ and let $X_{(1,\|\cdot\|)}, X_{(2,\|\cdot\|)}, \ldots, X_{(n,\|\cdot\|)}$ be the points in the sample in order of distance from X. If $m \geq 1$ is a constant and the weight function $W_{ni}(X)$ is a function of the query X, the sample points $X_1, X_2, \ldots, X_n$, and an independent random variable V, has support that is a subset of $X_{(1,\|\cdot\|)}, X_{(2,\|\cdot\|)}, \ldots, X_{(mk,\|\cdot\|)}$ (that is, it is nonzero only on the nearest mk points to X in the $\|\cdot\|$ norm) and $\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(X)\right] \to 0$, then the weight based classifier (of the form in equation (2.3)) is universally consistent.*

Proof: We see that the weights W_{ni} depend only on the query X, the sample points $X_1, X_2, \ldots, X_n$, and an independent random variable V, so we can apply Stone's theorem. We check the conditions of Stone's theorem:

1. For the first condition, we define c cones and mark the mk nearest neighbours of the query X in each cone, as in Lemma 2.3.8. By the argument in Lemma 2.3.8 (replacing the k nearest neighbour by mk nearest neighbour) we find that

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[W_{ni}(X) f(X_i)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \frac{1}{k} \mathbb{1}_{\{X_i \text{ is a } mk\text{-nearest neighbour of } X \text{ in the } \|\cdot\| \text{ norm among } X_1, X_2, \ldots, X_n\}} f(X_i)\right] \\
&= \frac{1}{k} \mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}_{\{X \text{ is a } mk\text{-nearest neighbour of } X_i \text{ in the } \|\cdot\| \text{ norm among } X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}} f(X)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{k}\mathbb{E}\left[f(X)\sum_{i=1}^{n}\mathbb{1}_{\{X \text{ is a } mk\text{-nearest neighbour of } X_i \text{ in the } \|\cdot\| \text{ norm among } X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\}}\right] \\
&\leq \frac{1}{k}\mathbb{E}\left[f(X)\sum_{i=1}^{n}\mathbb{1}_{\{X_i \text{ is marked}\}}\right] \\
&\leq \frac{m}{k}\mathbb{E}\left[f(X)(mk)c\right] \\
&= cm\mathbb{E}\left[f(X)\right].
\end{aligned}
$$

2. We see that for any fixed $a > 0$, $\mathbb{P}(\left\|X - X_{(mk,\|\cdot\|)}\right\| > a) \to 0$ as $n \to \infty$ by Lemma 2.3.4, and hence the second condition follows (as in Lemma 2.3.5, with k replaced by mk, since $k/n \to 0$ as $n \to \infty$, $mk/n \to 0$ as $n \to \infty$).

3. The third condition holds by assumption.

∎

Theorem 5.2.1 remains true if we replace the fixed norm $\|\cdot\|$ with a sequence $(\rho_n)_{n=1}^{\infty}$ of either random norms on $\mathbb{R}^d$ from a bounded family (from Theorem 3.2.5) or of random uniformly locally Lipschitz distances (from Theorem 4.1.9) such that the sequence is chosen independently of the query and of the sample. We now prove this for the most general case we have seen (a sequence of random uniformly locally Lipschitz distances, independent of the sample and the query).

Theorem 5.2.2. *Let $(\theta_n)_{n=1}^{\infty}$ be a sequence of random functions from a family of functions $\mathbb{R}^d$ to $\mathbb{R}$ satisfying the conditions of Theorem 4.1.9 (independent of the sample and the query) and let $X_{(1,\|\cdot\|)}, X_{(2,\|\cdot\|)}, \ldots, X_{(n,\|\cdot\|)}$ be defined as usual. If $m \geq 1$ is a constant and the weight function $W_{ni}(X)$ is a function of the query X, the sample points $X_1, X_2, \ldots, X_n$, and an independent random variable V, has support that is a subset of $X_{(1,\theta_n)}, X_{(2,\theta_n)}, \ldots, X_{(mk,\theta_{mk})}$ (that is, it is nonzero only on the nearest mk points to X in the θ_n distance) and $\mathbb{E}\left[\max_{1\leq i\leq n} W_{ni}(X)\right] \to 0$, then the weight based classifier (of the form in equation (2.3)) is universally consistent.*

Proof: We see that the weights W_{ni} depend only on the query X, the sample points $X_1, X_2, \ldots, X_n$, and an independent random variable V, so we can apply Stone's theorem. We check the conditions of Stone's theorem:

1. For the first condition, we proceed as in 4.1.8, we define c subsets in the same way as in the proof of the lemma and mark the mk nearest neighbours of the query X in each subsets. In the same manner as the proof of the lemma (replacing k with mk, which does not change anything else because $mk/n \to 0$ as $n \to \infty$ since $k/n \to 0$ as $n \to \infty$ and m is a fixed constant), we have

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right] \leq c\mathbb{E}\left[f(X)\right] + \epsilon_n.$$

2. For the second condition, we have

$$\mathbb{E}\left[\sum_{i=1}^{n} W_{ni}(X) \mathbb{1}_{\{\|X_i - X\| > a\}}\right] \leq \mathbb{P}\left(\left\|X_{(k,\|\cdot\|)} - X\right\| > \frac{\beta}{2\alpha} \min\{a, \delta\}\right).$$

   by Lemma 4.1.5, and since $mk/n \to 0$ as $n \to \infty$, by Lemma 2.3.4 the probability on the right hand side goes to zero as n approaches infinity, so the expectation on the left hand side (which by definition is nonnegative) goes to zero.

3. The third condition holds by assumption.

∎

Suppose we take either a fixed norm or an independent sequence of Lipschitz distance for k-NN, and we take the mk nearest points to the query at each step, with $m \geq 1$ a fixed constant. The above results allow us to adaptively pick a distance for k-NN based on the sample points (but not the sample labels), the query, and an independent random variable, with those mk nearest points to the query (that

is, we only consider those mk points in k-NN and ignore the rest). By Theorem 5.2.2 (or Theorem 5.2.1 for a fixed norm) we have that this results in a universally consistent classifier, since $W_{ni}(X)$ is nonzero only on the mk nearest neighbours in the θ_n distance (or $\|\cdot\|$ distance) and the maximum of the weights is $1/k$ which goes to zero as $n \to \infty$ (since $k \to \infty$ as $n \to \infty$).

To generate random variables independent of the sample and the query that are useful for us, we can independently randomly split the original sample into two smaller samples (with a fixed proportion of the points going into each sample). One of the subsamples becomes the sample used directly for classification (the weights can be nonzero for these points), and the other subsample becomes a set of points independent from the sample and the query which we can use for selecting the distance for k-NN. We then have points with their label which we can use in an optimization procedure for the distance while preserving universal consistency (since they are independent of the sample used for classification and the query, they become part of the independent random variable in the weight function). For instance, if we have an original sample of $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{2n}, Y_{2n})$ containing $2n$ points, we can split it into a sample of size n for the k-NN classifier and another disjoint set of points of size n which we use to find the distance for k-NN. For the global θ_n distance to find the mk nearest points, we use just those n points, for the local distance for the k nearest of those points, we can also use the query and sample points (but not labels) as well as those independent points.

Chapter 6

Datasets and Experimental Results

In this chapter we discuss classifying datasets based on the techniques described above. We consider both using a fixed distance for k-NN for the entire dataset and *locally chosen* distances, where we select the distance for k-NN based on the query and the sample. We first discuss some optimization methods used for selecting good distances for k-NN and a method for assessing the accuracy of our classification (where we run many trials for an accuracy and stable estimate). We then give examples showing classification accuracy improvements (compared to k-NN with the Euclidean norm) for a variety of datasets.

6.1 Methodology

We would like to find parameters (such as the p for the ℓ^p norm, the entries of a matrix, etc.) to achieve as high of a classification accuracy as possible. One approach is to try to maximize the classification accuracy on the training set. This has the disadvantage that the classification accuracy (for an empirical sample) is a step function, that is constant with sudden jumps. Optimization of such functions (especially when there are many parameters) can be very difficult. Optimizing the classification accuracy

for locally chosen distances on the training set produced poor results on the testing set for our datasets.

Another approach is to look at the correlation between distance and whether or not the points have the same label. We recall that the *Pearson's correlation coefficient* (often denoted $\rho_{X,Y}$) between two random variables X, Y is defined by

$$\mathrm{Corr}(X,Y) = \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{6.1}$$

It can be shown that $-1 \leq \mathrm{Corr}(X,Y) \leq 1$ always, with a correlation of 1 denoting a perfect positive linear relationship and a correlation of -1 denoting a perfect negative linear relationship between X and Y.[26] Suppose we select a point in the training set, which we call a training query. If we let X be the distance between the training query and other points in the dataset and Y be whether or not the label of the training query agrees with the label of the other point, a negative correlation between X and Y indicates that as we move away from the training query, we are more likely to find points with a different label, and if we move towards the training query we are more likely to find points with the same label. We then attempt to find a distance such that this correlation is as close to -1 as possible. To optimize over a family of distances to find a good locally chosen distance (for a query in the testing set), we can proceed as follows: we first take the k_1 and k_2 nearest neighbours of the query (in a fixed norm), with $k_2 \geq k_1$; we then take each of the k_1 nearest neighbours as a training query and find the correlation with the k_2 nearest neighbours described above, we minimize the mean correlation for each of the points in k_1 as a training query, the parameters we found are the ones used for the locally chosen distance for this testing set query. In this approach, we only consider the k_2 nearest points to the query in the correlation, and so only points in that neighbourhood affect the locally chosen distance. The correlation has the advantage that for our family of distances it is continuous (and in some cases differentiable), so optimization is far easier.

For optimizing our parameters, we use the R function `optim` to do the optimiza-

tion, which implements a variety of optimization methods. We describe some of these below. The problems are traditionally formulated as minimization problems, we can easily convert them into to a maximization problems by multiplying the function by -1 (in R, this can be done with the `fnscale=-1` option). The following brief descriptions are based on the book [19] and the R documentation [22].

The default method is the *Nelder-Mead method* of optimization (which is also known as the *Downhill Simplex method* or the *amoeba method*). It is well suited to optimization problems in multiple dimensions and uses only function values, without assuming the existence of derivatives. A *simplex* in $\mathbb{R}^d$ is a d-dimensional generalization of a tetrahedron (a triangle in $\mathbb{R}^2$, a tetrahedron in $\mathbb{R}^3$, and so on). In the Nelder-Mead method, we move the simplex using expansion, reflection, reflection and expansion, contraction, or multiple contraction at each step as necessary towards the minimum. The simplex thereby moves downhill towards a minimum. This is a robust method, but is relatively slow.

The *Broyden-Fletcher-Goldfarb-Shanno* (BFGS) method uses function values and gradients to build a picture of the surface to optimize. BFGS is a quasi-Newton method that computes an approximation to the Hessian matrix, which is updated at each step.

The *Conjugate Gradient* (CG) method (originally developed by Fletcher and Reeves (1964), there are options available to use modified versions) also uses gradients. The CG method does not compute a Hessian matrix approximation and so is better suited for large optimization problems, however, it tends to be more fragile than BFGS.

In the *Simulated Annealing* method, we try to find the global minimum of the function. Starting with some point, we generate new points around the current point randomly, based on a temperature parameter (using a Gaussian Markov kernel by default, in the R `optim` implementation). We sometimes accept points worse than the existing point, to avoid getting stuck in a local minimum (how often we do so depends

on the temperature). We then slowly decrease the temperature with time. This method uses only function values and works well for functions with noisy surfaces, but is relatively slow and is sensitive to the control parameters (like temperature) passed to it.

For each dataset described below, we find the accuracy by splitting the dataset into training and testing sets and finding the accuracy on the testing set. We repeat this cross-validation procedure multiple times, the exact details are described separately for each dataset. We have found (using Q-Q plots and the Shapiro-Wilk normality tests) that the empirical classification accuracies appear to have an approximate normal distribution. This means that we can find the $1-\alpha$ confidence interval for the mean accuracy using the formula (based on the Student's t distribution)

$$\overline{x} \pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}} \tag{6.2}$$

where $\overline{x}$ is the sample accuracy, n is the sample size, and s is the standard deviation of the observed data (using Bessel's correction of using the denominator $n-1$ instead of n to yield an unbiased estimator). [10]

When we apply k-NN, we determine the value of k by an empirical optimization procedure, in which we test all values of k up to a certain threshold (for instance, all k between 1 and 20) and use the value of k that produces the best accuracy. We apply this empirical optimization procedure for k on the training set (that is, we split the training set, and find the best value of k on this set), and we use the optimal k we found as the value of k for k-NN in the actual classification on the testing set. For the locally chosen distances, we select points near the point we found (near in the original distance) and find the optimal value of k for these points, which we use as the value of k for classifying the query with the locally chosen distance.

6.2 Experimental Results

6.2.1 Computer Generated Polynomials Dataset

We first look at the classification accuracy for a computer generated dataset we created. Each data point is a random vector in $\mathbb{R}^7$ that is uniformly distributed on $[0, 1]$ in each coordinate. For a data point $\boldsymbol{x} = (x_1, x_2, \ldots, x_7)$, we define $t = 2x_1$, and we evaluate the polynomials $p_1(t) = x_2t + x_3t^2 + x_4t^3$ and $p_2(t) = x_5t + x_6t^2 + x_7t^3$. If $p_1(t) > p_2(t)$, we assign label one, otherwise we assign label zero. We generate 50 such datasets, each generated independently and containing 500 points in the training set and another 500 points in the testing set. We then find the classification accuracy with k-NN in each case, trying various ℓ^p norms/quasinorms and locally chosen distances (in which we first multiply the data by a matrix and then apply either an ℓ^p norm or a polynomial as our distance, with these parameters being determined locally based on the labelled sample and the query with the above procedure). A table of the results we obtained is in Table 6.1, a box-and-whiskers plot is in Figure 6.1, and a plot of the 95 % confidence intervals is in 6.2. From this we see that the locally chosen ℓ^p distance with matrix produces the best results, followed by the fixed ℓ^p norms (with the accuracy being better for p significantly larger than one, and being roughly constant for such p). The ℓ^p quasinorms (with $0 < p < 1$) perform poorly on this dataset.

6.2.2 Face Recognition Dataset

We compare various norms on the CDMC2013 face recognition task (from [32]). We have a dataset consisting of 864 image vectors with 2576 dimensions, with 216 classes (with each class repeated exactly four times in the dataset). We use stratified sampling where for each class we select one image vector to be in our testing set and the other three to be in our training set. We first apply Principal Component Analysis

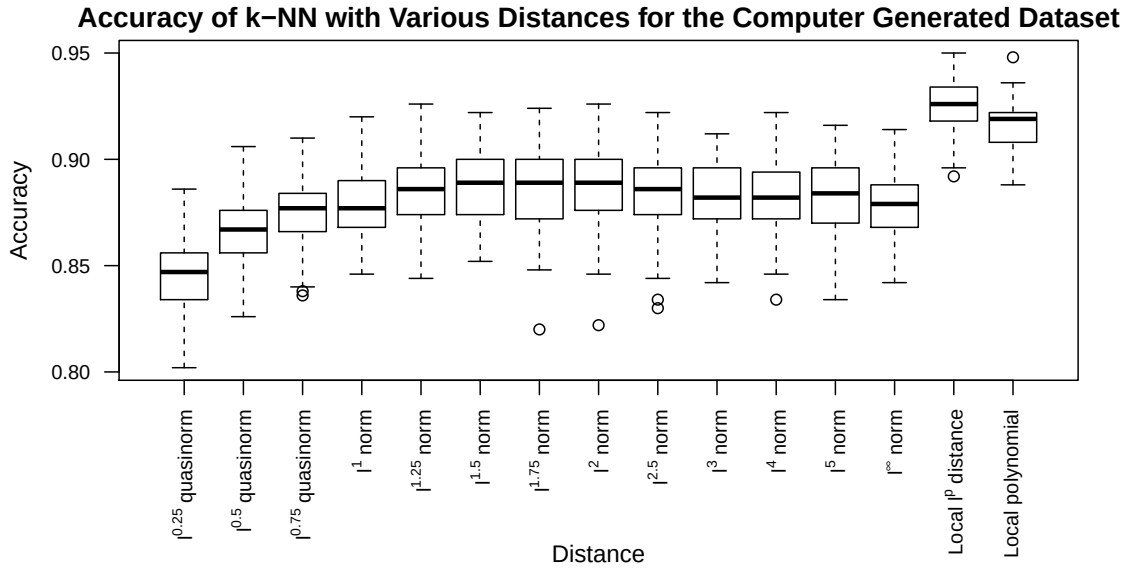

Figure 6.1: Box-and-whiskers plot of the accuracy of k-NN with various ℓ^p norms and locally chosen distances for the computer generated dataset.

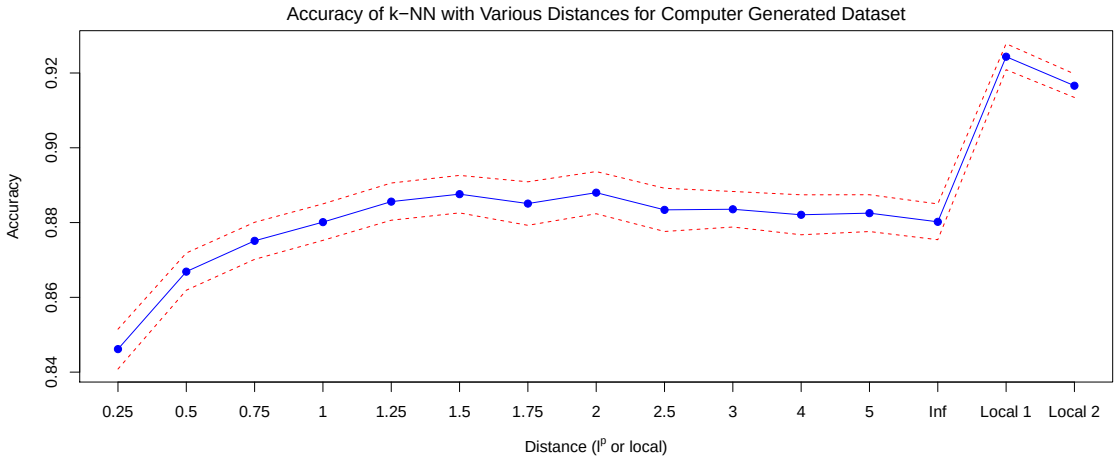

Figure 6.2: Plot of the accuracy of k-NN with various ℓ^p norms, quasinorms, and locally chosen distances for computer generated dataset (showing 95 % confidence intervals around the mean result, and data points). Here local 1 is locally chosen ℓ^p distance with matrix, local 2 is locally chosen polynomial with matrix.

Distance	Mean Accuracy	95% Confidence Interval
$\ell^{0.25}$ quasinorm	0.84616	$[0.8408282, 0.8514918]$
$\ell^{0.5}$ quasinorm	0.86688	$[0.8619008, 0.8718592]$
$\ell^{0.75}$ quasinorm	0.87512	$[0.870166, 0.880074]$
ℓ^1 norm	0.88012	$[0.8752407, 0.8849993]$
$\ell^{1.25}$ norm	0.8856	$[0.8806249, 0.8905751]$
$\ell^{1.5}$ norm	0.8876	$[0.8825867, 0.8926133]$
$\ell^{1.75}$ norm	0.88508	$[0.8792594, 0.8909006]$
ℓ^2 norm	0.888	$[0.882358, 0.893642]$
$\ell^{2.5}$ norm	0.8834	$[0.877618, 0.889182]$
ℓ^3 norm	0.88356	$[0.8788055, 0.8883145]$
ℓ^4 norm	0.88208	$[0.8767357, 0.8874243]$
ℓ^5 norm	0.88252	$[0.8775985, 0.8874415]$
ℓ^∞ norm	0.8802	$[0.8754124, 0.8849876]$
Local Distance with ℓ^p norm and matrix	0.92436	$[0.9208851, 0.9278349]$
Local Distance with degree 5 polynomial and matrix	0.9166	$[0.9134567, 0.9197433]$

Table 6.1: The mean accuracy and confidence intervals for k-NN applied to the computer generated dataset with various ℓ^p norms and local distances.

ℓ^p norm	Mean Accuracy	99% Confidence Interval for this Sample Dataset
1	0.6341667	$[0.6246978, 0.6436356]$
1.25	0.7075926	$[0.6967336, 0.7184516]$
1.5	0.6585185	$[0.6469469, 0.6700901]$
1.75	0.5660185	$[0.5534964, 0.5785407]$
2	0.4932407	$[0.4808274, 0.5056541]$

Table 6.2: The mean accuracy and confidence intervals for k-NN applied to the Face Recognition dataset with various ℓ^p norms.

(PCA)[33], reduce the dimension to 864, and calculate median centroids (that is, we calculate the median of each feature for all image vectors of the same class in the training set). Following this, we apply k-NN with $k = 1$ to the median centroids. We find that the $\ell^{1.25}$ norm is optimal on this dataset, with norms near the $\ell^{1.25}$ norm having similar performance and norms further away having worse performance.

When we apply k-NN with various ℓ^p norms to this dataset, we obtain the results in the box-and-whiskers plot 6.3. We have also tested various other norms which where found to perform far worse, in particular the quasinorms with $p < 1$ have been found to perform very poorly. In Table 6.2, we give some mean accuracies for various ℓ^p norms for this dataset (the confidence intervals are from sampling training and testing sets from this particular dataset, with 4 possibilities for each class to be chosen for the training set, this should not be interpreted as 99 % confidence intervals for the actual misclassification error for the underlying distribution for the data). Figure 6.3 gives a box-and-whiskers plot of the accuracies for many ℓ^p norms with $1 \leq p \leq 2$. The results for p outside this range were poor, with performance dropping off for $p < 1$ or $p > 2$. We can see from this that the $\ell^{1.25}$ norm (and ℓ^p norms with p near 1.25) gives the best result here.

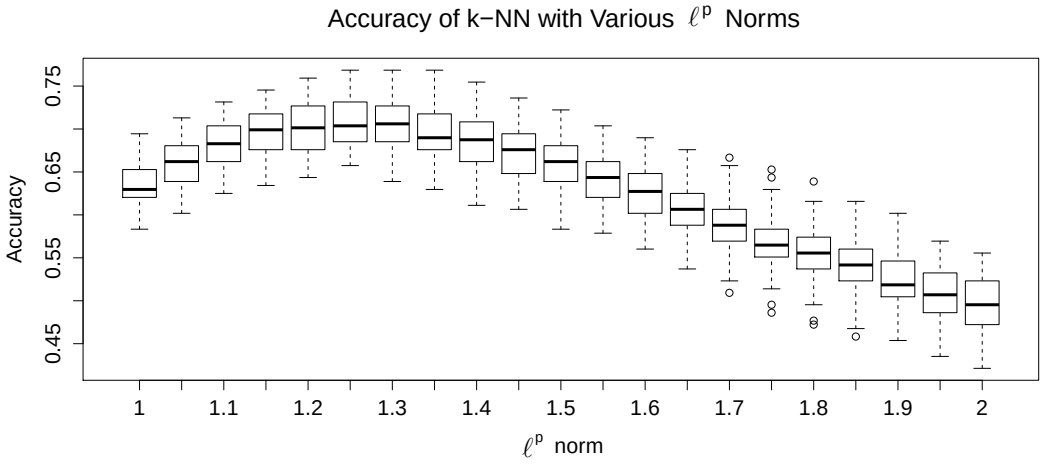

Figure 6.3: Box-and-whiskers plot of the accuracy of k-NN with various ℓ^p norms for the Face Recognition dataset.

6.2.3 Forest Cover Dataset

We now look at a subset of a forest cover dataset from [31], which was the subject of a Kaggle competition. We do not include the categorical data from the original dataset, only the numerical part (hence our results below are not comparable to the Kaggle competition, in practice we would use this in conjunction with a classifier for the categorical part as the categorical part is very important to achieve a high classification accuracy, classification accuracies much higher than ours are possible using the categorical part alone). There are 10 numerical columns, we fit all of them to the interval $[0, 1]$.

The original dataset contains 15120 rows. We take 50 random samples, each of them containing 2000 rows for the training set and 1000 rows for the testing set (with the training and testing sets disjoint in each case). We then find the classification accuracy with k-NN in each case, trying various ℓ^p norms/quasinorms and locally chosen distances (for the locally chosen distances, we first multiply the data by a matrix and then apply either an ℓ^p norm or a polynomial as our distance, with these

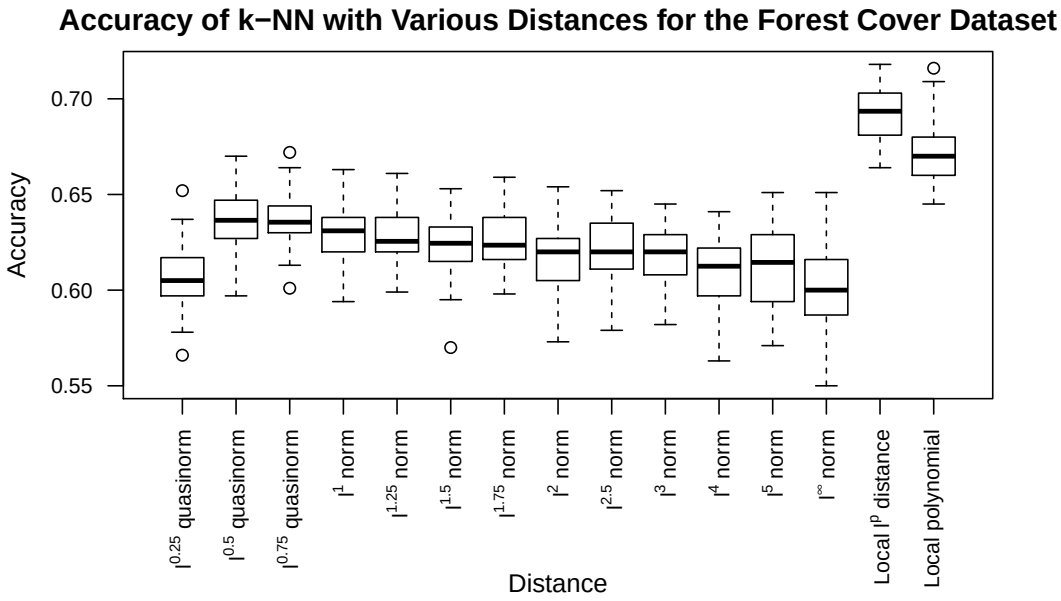

Figure 6.4: Box-and-whiskers plot of the accuracy of k-NN with various ℓ^p norms and locally chosen distances for the Forest Cover dataset.

parameters being determined locally based on the labelled sample and the query with the above procedure). In particular, we have tried the ℓ^p norms and quasinorms with p being $0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4, 5, \infty$, and a couple of locally chosen distances, one consisting of optimizing over matrices and ℓ^p norms, the other consisting of optimizing over matrices and polynomials of degree 5. The mean accuracies obtained (with 95 % confidence intervals) are given in Table 6.3, a box-and-whiskers plot is shown in Figure 6.4, and a plot of the accuracies with 95 % confidence intervals is in Figure 6.5. We see that both locally chosen distances deliver superior performance to any fixed norm that was tested, with local ℓ^p norms with matrices being better than local polynomials with matrices. For fixed ℓ^p norms/quasinorms, the accuracy seems to be highest around 0.5 and 0.75, with smaller p performing much worse and the accuracy dropping off as p increases beyond 0.75.

Distance	Mean Accuracy	95% Confidence Interval
$\ell^{0.25}$ quasinorm	0.60538	$[0.6005078, 0.6102522]$
$\ell^{0.5}$ quasinorm	0.63666	$[0.6323007, 0.6410193]$
$\ell^{0.75}$ quasinorm	0.63652	$[0.6327927, 0.6402473]$
ℓ^1 norm	0.6309	$[0.6270883, 0.6347117]$
$\ell^{1.25}$ norm	0.62892	$[0.6251664, 0.6326736]$
$\ell^{1.5}$ norm	0.6229	$[0.6187195, 0.6270805]$
$\ell^{1.75}$ norm	0.6252	$[0.6210497, 0.6293503]$
ℓ^2 norm	0.61824	$[0.6135345, 0.6229455]$
$\ell^{2.5}$ norm	0.62046	$[0.615441, 0.625479]$
ℓ^3 norm	0.61784	$[0.6072968, 0.6177832]$
ℓ^4 norm	0.60872	$[0.6033483, 0.6140917]$
ℓ^5 norm	0.61254	$[0.6072968, 0.6177832]$
ℓ^∞ norm	0.59996	$[0.5939458, 0.6059742]$
Local distance with ℓ^p norm and matrix	0.69272	$[0.6890909, 0.6963491]$
Local distance with degree 5 polynomial and matrix	0.67222	$[0.6675009, 0.6769391]$

Table 6.3: The mean accuracy and confidence intervals for k-NN applied to the Forest Cover dataset with various ℓ^p norms and locally chosen distances.

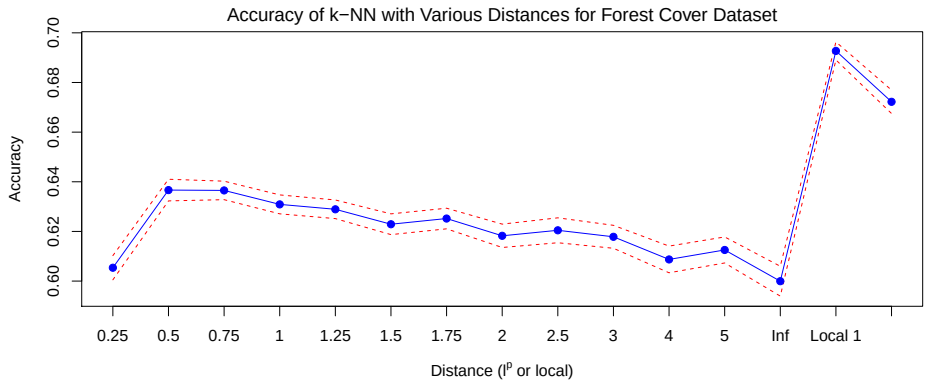

Figure 6.5: Plot of the accuracy of k-NN with various ℓ^p norms, quasinorms, and locally chosen distances for Forest Cover dataset (showing 95 % confidence intervals around the mean result, and data points). Here local 1 is locally chosen ℓ^p distance with matrix, local 2 is locally chosen polynomial with matrix.

6.2.4 Higgs Boson Dataset

The ATLAS Higgs Boson dataset ([29]) contains 29 numeric data columns and a response column with two possible states (namely, whether an event is a Higgs Boson event or is background noise). There are other columns, but they are not relevant for us (they include an importance weight for an alternative classification performance measure). We remove data columns with missing values to obtain 18 data columns. The original dataset contains 818238 rows. We then randomly generate 500 independent random subsets of the original dataset, each containing 5000 training rows and 5000 testing rows (with the training and testing sets being disjoint). Before applying k-NN, we fit the columns fitted to the interval $[0, 1]$.

We then find the classification accuracy with k-NN in each case, trying various ℓ^p norms/quasinorms. In particular, we have tried the ℓ^p norms and quasinorms with p being $0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 4, 5, \infty$. The mean accuracies obtained (with 95 % confidence intervals) are given in Table 6.4, a box-and-whiskers

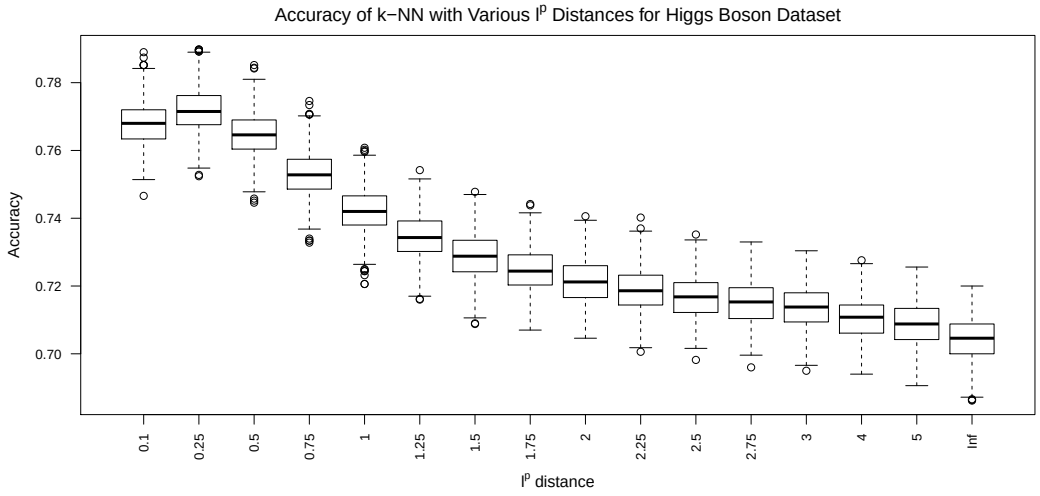

Figure 6.6: Box-and-whiskers plot of the accuracy of k-NN with various ℓ^p norms and quasinorms for the Higgs Boson dataset.

plot is shown in Figure 6.6. We see that the ℓ^p quasinorms with $0 < p < 1$ perform better than the ℓ^p norms with $p \geq 1$, with $p = 1/4$ (the $\ell^{1/4}$ quasinorm) resulting in the highest classification accuracy. We see that after $p = 1/4$, the classification accuracy decreases as p increases. The improvement in accuracy in using the $\ell^{1/4}$ quasinorm is very significant, the mean accuracy with the $\ell^{1/4}$ quasinorm is approximately 77.2 % while for the Euclidean norm it is approximately 72.1 % (so the quasinorm performs approximately 5 % better than the Euclidean norm).

Distance	Mean Accuracy	95% Confidence Interval
$\ell^{0.1}$	0.7678608	$[0.7672895, 0.7684321]$
$\ell^{0.25}$	0.7718852	$[0.7713035, 0.7724669]$
$\ell^{0.5}$	0.7646396	$[0.7640515, 0.7652277]$
$\ell^{0.75}$	0.7529876	$[0.7523944, 0.7535808]$
ℓ^{1}	0.7423464	$[0.7417668, 0.742926]$
$\ell^{1.25}$	0.7345884	$[0.7340186, 0.7351582]$
$\ell^{1.5}$	0.7289776	$[0.7283997, 0.7295555]$
$\ell^{1.75}$	0.724774	$[0.7242062, 0.7253418]$
ℓ^{2}	0.7214408	$[0.720877, 0.7220046]$
$\ell^{2.25}$	0.718962	$[0.718397, 0.719527]$
$\ell^{2.5}$	0.716842	$[0.7162883, 0.7173957]$
$\ell^{2.75}$	0.7152068	$[0.7146595, 0.7157541]$
ℓ^{3}	0.7138832	$[0.713334, 0.7144324]$
ℓ^{4}	0.7104716	$[0.7099282, 0.711015]$
ℓ^{5}	0.708712	$[0.7081656, 0.7092584]$
ℓ^{∞}	0.7042	$[0.7036437, 0.7047563]$

Table 6.4: The mean accuracy and confidence intervals for k-NN applied to the Higgs Boson dataset with various ℓ^p norms.

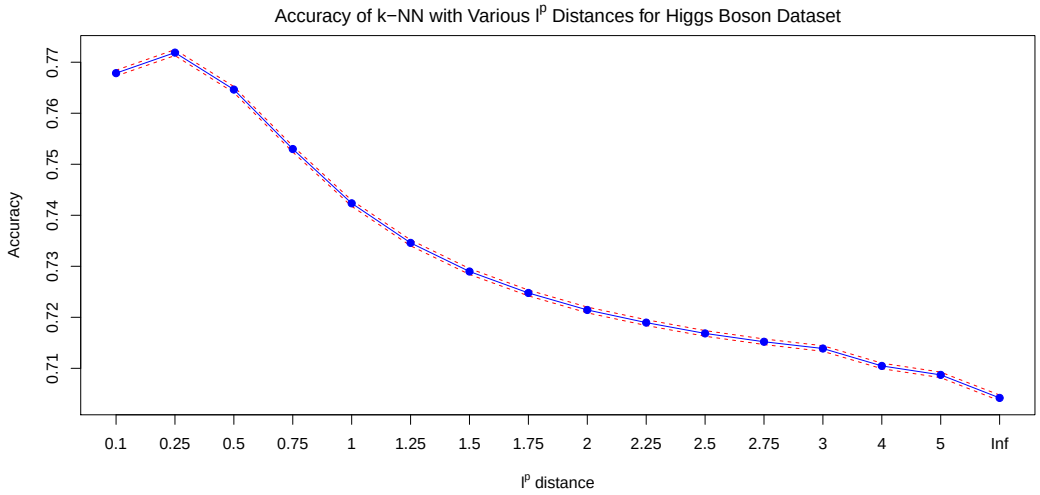

Figure 6.7: Plot of the accuracy of k-NN with various ℓ^p norms and quasi-norms for the Higgs Boson dataset (showing 95 % confidence intervals around the mean result, and data points).

Chapter 7

Future Prospects

There are many possible future projects based on this work. Here is an outline of some possibilities:

- We would like to extend our result in Theorem 3.2.5 to be able to handle norms chosen at each step based on the sample points $X_1, X_2, \ldots, X_n$ and possibly the query X (and similarly for Theorem 4.1.9 with the uniformly locally Lipschitz family). In [14], a similar result is claimed in Theorem 26.3 (where we multiply the data by a matrix that is a function of the sample points $X_1, X_2, \ldots, X_n$ and then apply the Euclidean norm), however the proof provided is incomplete (part of it being incorrect), as we have shown above. We would like to recover that result, which should hold for more general families of norms.

- There is a classification algorithm (described in [30]) called the *large margin nearest neighbour* (*LMNN*), in which we learn a positive semi-definite matrix M used to construct a pseudometric of the form $\rho(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^\intercal M(\boldsymbol{x} - \boldsymbol{y})$ for k-NN (a *pseudometric* is similar to a metric, but can take value zero for distinct points). The procedure for learning the matrix M depends on the query and the labelled sample. LMNN has been found to produce good results for classifying various datasets We would like to determine if LMNN (or a similar algorithm)

is universally consistent.

- Here, we have proven the universal consistency of k-NN based on Stone's theorem, and have applied this to various modifications of the classical k-NN classifier (such as using a sequence of norms or of Lipschitz functions). There is an alternative proof for the universal consistency of k-NN based on the Lebesgue-Besicovitch Differentiation Theorem, originally given by Luc Devroye in [34] and discussed further by Frédéric Cérou and Arnaud Guyader in the paper [24]. Priess has shown (in [25]) that the conclusion of the differentiation theorem is equivalent to the σ-finite dimensionality of a metric space. We would like to extend this result to sequences of norms (instead of a single fixed norm, similar to our result) and investigate if it can lead to a universal consistency proof with quasinorms and various other distances.

- We would like to improve our optimization methods, so we can more accurately and rapidly optimize over a class of distances for k-NN. We have found some like the correlation method discussed above, we would like to find others.

- We have assumed a bounded family of norms (or uniformly locally Lipschitz distances). This was a necessary assumption as we saw for the sequences of norms given by equations (3.24) and (3.25). We notice that such a sequence is extremely unlikely to be picked by an optimizer optimizing over the family of norms for that distribution (indeed, the probability approaches zero as n approaches infinity for our distribution). We would like to determine if we can remove this assumption provided we follow an optimization procedure instead of picking an arbitrary (possibly bad) sequence. We may find a solution to this problem if we develop a theory of capacity for norms for k-NN, similar to VC dimension (Vapnik-Chervonenkis dimension) and more generally metric entropy.

Bibliography

[1] Pestov, V., *Is the k-NN classifier in high dimensions affected by the curse of dimensionality?*, Computers and Mathematics with Applications (2012), doi:10.1016/j.camwa.2012.09.011.

[2] Stone, C., *Consistent Nonparametric Regression*, Annals of Statistics, 1977.

[3] Kechris, A., *Classical Descriptive Set Theory*, Springer-Verlag, 1995.

[4] Stephen H. Friedberg, Arnold J. Insel, and Lawrence E. Spence, *Linear Algebra*, Pearson Prentice Hall, 2003.

[5] Ben Noble, James W. Daniel, *Applied Linear Algebra* (Third Edition), Prentice Hall, 1987.

[6] Robert S. Strichartz, *The Way of Analysis*, Jones and Bartlett Learning, 2000.

[7] Gerald B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Second Edition (Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts), Wiley, 1999.

[8] Andrew M. Bruckner, Judith B. Bruckner, Brian S. Thomson, *Real Analysis*, Second Edition, ClassicalRealAnalysis.com, 2008, xiv 656 pp. [ISBN 1434844129].

[9] David McDonald, *Elements of Applied Probability for Engineering, Mathematics, and Systems Science*, World Scientfic, 2004.

[10] Robert V. Hogg and Elliot A. Tanis, *Probability and Statistical Inference*, Eighth Edition, Pearson Prentice Hall, 2010.

[11] W. Keith Nicholson, *Introduction to Abstract Algebra*, Third Edition, Wiley-Interscience, 2006.

[12] Graeme Cohen, *A Course in Modern Analysis and its Applications*, Cambridge University Press, 2003.

[13] Dudley, R. M., *Real Analysis and Probability* (Cambridge Studies in Advanced Mathematics), Cambridge University Press, 2004.

[14] Luc Devroye, László Györfi, Gábor Lugosi, *A Probabilistic Theory of Pattern Recognition* (Stochastic Modelling and Applied Probability), Springer, 1996.

[15] E. Hairer and G. Wanner, *Analysis by Its History*, (Undergraduate Texts in Mathematics: Readings in Mathematics), Springer, 1996.

[16] William J. Stewart, *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*, Princeton University Press, 2009.

[17] Elias M. Stein and Rami Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces* (Princeton Lectures in Analysis), Princeton University Press, 2005.

[18] Hubert Haoyang Duan, *Applying Supervised Learning Algorithms and a New Feature Selection Method to Predict Coronary Artery Disease*, University of Ottawa, 2013.

[19] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery, *Numerical Recipes in C (The Art of Scientific Computing)*, Second Edition, Cambridge University Press, 1992.

[20] Cover, T.; Hart, P., *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, vol.13, no.1, pp.21, 27, January 1967, doi: 10.1109/TIT.1967.1053964.

[21] Stan Hatko, *Borel Isomorphic Dimensionality Reduction of Data and Supervised Learning*, 2013, arXiv:1307.8333 [stat.ML].

[22] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[23] k-NN illustration from the page `https://en.wikipedia.org/wiki/File:KnnClassification.svg`.

[24] Frédéric Crou and Arnaud Guyader (2006). *Nearest neighbor classification in infinite dimension*. ESAIM: Probability and Statistics, 10, pp 340-355. doi:10.1051/ps:2006014.

[25] D. Preiss, *Dimension of metrics and differentiation of measures*. In General Topology and its Relations to Modern Analysis and Algebra V, pages 565-568, Heldermann Verlag, Berlin, 1983.

[26] George Casella and Roger L. Berger, *Statistical Inference*, Duxbury Advanced Series, 2002.

[27] Patrick Billingsley, *Probability and Measure*, Third Edition, Wiley Series in Probability and Mathematical Statistics, 1995.

[28] Andrew Pressley, *Elementary Differential Geometry*, Second Edition, Springer Undergraduate Mathematics Series, Springer-Verlag, 2012.

[29] ATLAS collaboration (2014). *Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014*. CERN Open Data Portal. DOI: 10.7483/OPENDATA.ATLAS.ZBP2.M5T8.

[30] Kilian Q. Weinberger and Lawrence K. Saul, *Distance Metric Learning for Large Margin Nearest Neighbor Classification*, Journal of Machine Learning Research 10 (2009) 207-244.

[31] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

[32] S. Pang and L. Zhu. (2013). DMLI multi-task face recognition dataset. Unitec Institute of Technology New Zealand, Decentralized Machine Learning Intelligence (DMLI).

[33] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (2012). *Foundations of Machine Learning*, The MIT Press.

[34] Luc Devroye (1981), *On the Almost Everywhere Convergence of Nonparametric Regression Function Estimates*, The Annals of Statistics, 1981, Vol. 9, No . 6, 1310-1319.

[35] Patrice Assouad and Thierry Quentin de Gromard (2006), *Recouvrements, derivation des mesures et dimensions*, Rev. Mat. Iberoamericana 22 (2006), no. 3, 893953.

[36] Gérard Biau, Luc Devroye and Gábor Lugosi (2008), *Consistency of Random Forests and Other Averaging Classifiers*, Journal of Machine Learning Research 9 (2008) 2015-2033.