Journal of Machine Learning Research 10 (2009) 657-698 Submitted 10/07; Revised 6/08; Published 3/09

Nearest Neighbor Clustering: A Baseline Method for Consistent Clustering with Arbitrary Objective Functions

Sébastien Bubeck SEBASTIEN.BUBECK@INRIA.FR
SequeL Project, INRIA Lille
40 avenue Halley,
59650 Villeneuve d'Ascq, France

Ulrike von Luxburg ULRIKE.LUXBURG@TUEBINGEN.MPG.DE
Max Planck Institute for Biological Cybernetics
Spemannstr. 38,
72076 Tübingen, Germany

Editor: Charles Elkan

Abstract

Clustering is often formulated as a discrete optimization problem. The objective is to find, among all partitions of the data set, the best one according to some quality measure. However, in the statistical setting where we assume that the finite data set has been sampled from some underlying space, the goal is not to find the best partition of the given sample, but to approximate the true partition of the underlying space. We argue that the discrete optimization approach usually does not achieve this goal, and instead can lead to inconsistency. We construct examples which provably have this behavior. As in the case of supervised learning, the cure is to restrict the size of the function classes under consideration. For appropriate "small" function classes we can prove very general consistency theorems for clustering optimization schemes. As one particular algorithm for clustering with a restricted function space we introduce "nearest neighbor clustering". Similar to the k-nearest neighbor classifier in supervised learning, this algorithm can be seen as a general baseline algorithm to minimize arbitrary clustering objective functions. We prove that it is statistically consistent for all commonly used clustering objective functions.

Keywords: clustering, minimizing objective functions, consistency

1. Introduction

Clustering is the problem of discovering "meaningful" groups in given data. In practice, the most common approach to clustering is to define a clustering quality function Q_n, and then construct an algorithm which is able to minimize (or maximize) Q_n. There exists a huge variety of clustering quality functions: the K-means objective function based on the distance of the data points to the cluster centers, graph cut based objective functions such as ratio cut or normalized cut, or various criteria based on some function of the within- and between-cluster similarities. Once a particular clustering quality function Q_n has been selected, the objective of clustering is stated as a discrete optimization problem. Given a data set $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ and a clustering quality function Q_n, the ideal clustering algorithm should take into account all possible partitions of the data set and output the one that minimizes Q_n. The implicit understanding is that the "best" clustering can be any partition out of the set of all possible partitions of the data set. The practical challenge is then

to construct an algorithm which is able to explicitly compute this "best" clustering by solving an optimization problem. We will call this approach the "discrete optimization approach to clustering".

Now let us look at clustering from the perspective of statistical learning theory. Here we assume that the finite data set has been sampled from an underlying data space $\mathcal{X}$ according to some probability measure $\mathbb{P}$. The ultimate goal in this setting is not to discover the best possible partition of the data set $\mathcal{X}_n$, but to learn the "true clustering" of the underlying space $\mathcal{X}$. While it is not obvious how this "true clustering" should be defined in a general setting (cf. von Luxburg and Ben-David, 2005), in an approach based on quality functions this is straightforward. We choose a clustering quality function Q on the set of partitions of the entire data space $\mathcal{X}$, and define the true clustering f^* to be the partition of $\mathcal{X}$ which minimizes Q. In a finite sample setting, the goal is now to approximate this true clustering as well as possible. To this end, we define an empirical quality function Q_n which can be evaluated based on the finite sample only, and construct the empirical clustering f_n as the minimizer of Q_n. In this setting, a very important property of a clustering algorithm is consistency: we require that $Q(f_n)$ converges to $Q(f^*)$ when $n \to \infty$. This strongly reminds of the standard approach in supervised classification, the empirical risk minimization approach. For this approach, the most important insight of statistical learning theory is that in order to be consistent, learning algorithms have to choose their functions from some "small" function space only. There are many ways how the size of a function space can be quantified. One of the easiest ways is to use shattering coefficients $s(\mathcal{F},n)$ (see Section 2 for details). A typical result in statistical learning theory is that a necessary condition for consistency is $\mathbb{E}\log s(\mathcal{F},n)/n \to 0$ (cf. Theorem 2.3 in Vapnik, 1995, Section 12.4 of Devroye et al., 1996). That is, the "number of functions" $s(\mathcal{F},n)$ in $\mathcal{F}$ must not grow exponentially in n, otherwise one cannot guarantee for consistency.

Stated like this, it becomes apparent that the two viewpoints described above are not compatible with each other. While the discrete optimization approach on any given sample attempts to find the best of all (exponentially many) partitions, statistical learning theory suggests to restrict the set of candidate partitions to have sub-exponential size. So from the statistical learning theory perspective, an algorithm which is considered ideal in the discrete optimization setting will not produce partitions which converge to the true clustering of the data space.

In practice, for most clustering objective functions and many data sets the discrete optimization approach cannot be performed perfectly as the corresponding optimization problem is NP hard. Instead, people resort to heuristics and accept suboptimal solutions. One approach is to use local optimization procedures potentially ending in local minima only. This is what happens in the K-means algorithm: even though the K-means problem for fixed K and fixed dimension is not NP hard, it is still too hard for being solved globally in practice. Another approach is to construct a relaxation of the original problem which can be solved efficiently (spectral clustering is an example for this). For such heuristics, in general one cannot guarantee how close the heuristic solution is to the finite sample optimum. This situation is clearly unsatisfactory: in general, we neither have guarantees on the finite sample behavior of the algorithm, nor on its statistical consistency in the limit.

The following alternative approach looks much more promising. Instead of attempting to solve the discrete optimization problem over the set of all partitions, and then resorting to relaxations due to the hardness of this problem, we turn the tables. Directly from the outset, we only consider candidate partitions in some restricted class $\mathcal{F}_n$ containing only polynomially many functions. Then the discrete optimization problem of minimizing Q_n over $\mathcal{F}_n$ is not NP hard—formally it can be solved in polynomially many steps by trying all candidates in $\mathcal{F}_n$. From a theoretical point of view this approach has the advantage that the resulting clustering algorithm has the potential of being

consistent. In addition, this approach also has advantages in practice: rather than dealing with uncontrolled relaxations of the original problem, we restrict the function class to some small subset $\mathcal{F}_n$ of "reasonable" partitions. Within this subset, we then have complete control over the solution of the optimization problem and can find the global optimum. Put another way, one can also interpret this approach as some controlled way to approximate a solution of the NP hard optimization problem on the finite sample, with the positive side effect of obeying the rules of statistical learning theory.

This is the approach we want to describe in this paper. In Section 2 we will first construct an example which demonstrates the inconsistency in the discrete optimization approach. Then we will state a general theorem which gives sufficient conditions for clustering optimization schemes to be consistent. We will see that the key point is to control the size of the function classes the clustering are selected from. In Section 3 we will then introduce an algorithm which is able to work with such a restricted function class. This algorithm is called nearest neighbor clustering, and in some sense it can be seen as a clustering-analogue to the well-known nearest neighbor classifier for classification. We prove that nearest neighbor clustering is consistent under minimal assumptions on the clustering quality functions Q_n and Q. Then we will apply nearest neighbor clustering to a large variety of clustering objective functions, such as the K-means objective function, normalized cut and ratio cut, the modularity objective function, or functions based on within-between cluster similarity ratios. For all these functions we will verify the consistency of nearest neighbor clustering in Section 4. Discussion of our results, also in the context of the related literature, can be found in Sections 5 and 6. The proofs of all our results are deferred to the appendix, as some of them are rather technical.

2. General (In)Consistency Results

In the rest of this paper, we consider a space $\mathcal{X}$ which is endowed with a probability measure $\mathbb{P}$. The task is to construct a clustering $f : \mathcal{X} \to \{1,\dots,K\}$ on this space, where K denotes the number of clusters to construct. We denote the space of all $\mathbb{P}$-measurable functions from $\mathcal{X}$ to $\{1,...,K\}$ by $\mathcal{H}$. Let $Q : \mathcal{H} \to \mathbb{R}^+$ denote a clustering quality function: for each clustering, it tells us "how good" a given clustering is. This quality function will usually depend on the probability measure $\mathbb{P}$. An optimal clustering, according to this objective function, is a clustering f^* which satisfies

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{F}} Q(f).$$

where $\mathcal{F} \subseteq \mathcal{H}$ is a fixed set of candidate clusterings. Now assume that $\mathbb{P}$ is unknown, but that we are given a finite sample $X_1,...,X_n \in \mathcal{X}$ which has been drawn i.i.d according to $\mathbb{P}$. Our goal is to use this sample to construct a clustering f_n which "approximates" an optimal clustering f^*. To this end, assume that $Q_n : \mathcal{H} \to \mathbb{R}^+$ is an estimator of Q which can be computed based on the finite sample only (that is, it does not involve any function evaluations $f(x)$ for $x \notin \{X_1,...,X_n\}$). We then consider the clustering

$$f_n \in \operatorname*{argmin}_{f \in \mathcal{F}_n} Q_n(f).$$

Here, $\mathcal{F}_n$ is a subset of $\mathcal{H}$, which might or might not be different from $\mathcal{F}$. The general question we are concerned with in this paper is the question of consistency: under which conditions do we know that $Q(f_n) \to Q(f^*)$?

Note that to avoid technical overload we will assume throughout this paper that all the minima (as in the definitions of f^* and f_n) exist and can be attained. If this is not the case, one can always go over to statements about functions which are ε-close to the corresponding infimum. We also will not discuss issues of measurability in this paper (readers interested in measurability issues for empirical processes are referred to Section 1 of van der Vaart and Wellner, 1996).

2.1 Inconsistency example

In the introduction we suggested that as in the supervised case, the size of the function class $\mathcal{F}_n$ might be the key to consistency of clustering. In particular, we argued that optimizing over the space of all measurable functions might lead to inconsistency. First of all, we would like to prove this statement by providing a example. This example will show that if we optimize a clustering objective function over a too large class of functions, the resulting clusterings are not consistent.

Example 1 (Inconsistency in general) *As data space we choose $\mathcal{X} = [0,1] \cup [2,3]$, and as probability measure $\mathbb{P}$ we simply use the normalized Lebesgue measure λ on $\mathcal{X}$. We define the following similarity function between points in $\mathcal{X}$:*

$$s(x,y) = \begin{cases} 1 & \text{if } x \in [0,1], y \in [0,1] \\ 1 & \text{if } x \in [2,3],\ y \in [2,3] \\ 0 & \text{otherwise.} \end{cases}$$

For simplicity, we consider the case where we want to construct $K = 2$ clusters called C_1 and C_2. Given a clustering function $f : \mathcal{X} \to \{0,1\}$ we call the clusters $C_1 := \{x \in \mathcal{X} \mid f(x) = 0\}$ and $C_2 := \{x \in \mathcal{X} \mid f(x) = 1\}$. As clustering quality function Q we use the between-cluster similarity (equivalent to cut*, see Section* 4.2 *for details):*

$$Q(f) = \int_{x \in C_1} \int_{y \in C_2} s(X,Y)\, d\mathbb{P}(X)\, d\mathbb{P}(Y).$$

As an estimator of Q we will use the function Q_n where the integrals are replaced by sums over the data points:

$$Q_n(f) = \frac{1}{n(n-1)} \sum_{i \in C_1} \sum_{j \in C_2} s(X_i, X_j).$$

As set $\mathcal{F}$ we choose the set of all measurable partitions on $\mathcal{X}$ (note that the same example also holds true when we only look at the set $\mathcal{F}$ of measurable partitions such that both clusters have a minimal mass ε for some $\varepsilon > 0$). For all $n \in \mathbb{N}$ we set $\mathcal{F}_n = \mathcal{F}$. Let $X_1, ..., X_n \in \mathcal{X}$ be our training data. Now define the functions

$$f^*(x) = \begin{cases} 0 & \text{if } x \in [0,1] \\ 1 & \text{if } x \in [2,3] \end{cases} \qquad \text{and} \qquad f_n(x) = \begin{cases} 0 & \text{if } x \in \{X_1, ..., X_n\} \cap [0,1] \\ 1 & \text{if } x \in [2,3] \\ 0 & \text{if } x \in [0, 0.5] \setminus \{X_1, ..., X_n\} \\ 1 & \text{if } x \in [0.5, 1] \setminus \{X_1, ..., X_n\} \end{cases}.$$

It is obvious that $Q(f^*) = 0$ and $Q_n(f_n) = 0$. As both Q and Q_n are non-negative, we can conclude $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} Q(f)$ and $f_n \in \operatorname{argmin}_{f \in \mathcal{F}} Q_n(f)$. It is also straightforward to compute $Q(f_n) = 1/16$ (independently of n). Hence, we have inconsistency: $1/16 = Q(f_n) \not\to Q(f^*) = 0$.

Note that the example is set up in a rather natural way. The data space contains two perfect clusters ($[0,1]$ and $[2,3]$) which are separated by a large margin. The similarity function is the ideal similarity function for this case, giving similarity 1 to points which are in the same cluster, and similarity 0 to points in different clusters. The function f^* is the correct clustering. The empirical clustering f_n, if restricted to the data points, reflects the correct clustering. It is just the "extension" of the empirical clustering to non-training points which leads to the inconsistency of f_n. Intuitively, the reason why this can happen is clear: the function space $\mathcal{F}$ does not exclude the unsuitable extension chosen in the example, the function overfits. This can happen because the function class is too large.

2.2 Main Result

Now we would like to present our first main theorem. It shows that if f_n is only picked out of a "small" function class $\mathcal{F}_n$, then we can guarantee consistency of clustering. Before stating the theorem we would like to recall the definition of the shattering coefficient in a K-class setting. For a function class $\mathcal{F} : \mathcal{X} \to \{1,\dots,K\}$ the shattering coefficient of size n is defined as

$$s(\mathcal{F},n) = \max_{x_1,\dots,x_n \in \mathcal{X}} |\{(f(x_1),\dots,f(x_n)) \mid f \in \mathcal{F}\}|.$$

To state our theorem, we will also require a pseudo-distance d between functions. A pseudo-distance is a dissimilarity function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ which is symmetric, satisfies the triangle inequality and the condition $f = g \implies d(f,g) = 0$, but not necessarily the condition $d(f,g) = 0 \implies f = g$. For distances between sets of functions we use the standard convention $d(\mathcal{F},\mathcal{G}) = \inf_{f\in\mathcal{F},g\in\mathcal{G}} d(f,g)$. Our theorem is as follows:

Theorem 1 (Consistency of a clustering optimizing scheme) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables which have been drawn i.i.d. according to some probability measure $\mathbb{P}$ on some set $\mathcal{X}$. Let $\mathcal{F}_n := \mathcal{F}_n(X_1,\dots,X_n) \subset \mathcal{H}$ be a sequence of function spaces, and $\mathcal{F} \subset \mathcal{H}$. Let $d : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ be a pseudo-distance defined on $\mathcal{H}$. Let $Q : \mathcal{H} \to \mathbb{R}^+$ be a clustering quality function, and $Q_n : \mathcal{H} \to \mathbb{R}^+$ an estimator of this function which can be computed based on the finite sample only. Finally let*

$$\widetilde{\mathcal{F}_n} := \bigcup_{X_1,\dots,X_n \in \mathbb{R}^d} \mathcal{F}_n.$$

Define the true and the empirical clusterings as

$$f^* \in \operatorname*{argmin}_{f\in\mathcal{F}} Q(f),$$
$$f_n \in \operatorname*{argmin}_{f\in\mathcal{F}_n} Q_n(f).$$

Assume that the following conditions are satisfied:

1. *$Q_n(f)$ is a consistent estimator of $Q(f)$ which converges sufficiently fast for all $f \in \widetilde{\mathcal{F}_n}$:*

$$\forall \varepsilon > 0,\ s(\widetilde{\mathcal{F}_n},2n) \sup_{f\in\widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| > \varepsilon) \to 0,$$

2. *$\mathcal{F}_n$ approximates $\mathcal{F}$ in the following sense:*

$$(i) \quad \forall f \in \mathcal{F}, d(f, \mathcal{F}_n) \to 0 \text{ in probability,}$$
$$(ii) \quad \mathbb{P}(f_n \notin \mathcal{F}) \to 0.$$

3. *Q is uniformly continuous with respect to the pseudo-distance d between $\mathcal{F}$ and $\widetilde{\mathcal{F}_n}$:*

$$\forall \varepsilon > 0 \; \exists \delta(\varepsilon) > 0 \text{ such that } \forall f \in \mathcal{F} \; \forall g \in \widetilde{\mathcal{F}_n} : \; d(f,g) \leq \delta(\varepsilon) \Rightarrow |Q(f) - Q(g)| \leq \varepsilon.$$

Then the optimization scheme is weakly consistent, that is $Q(f_n) \to Q(f^*)$ in probability.

This theorem states sufficient conditions for consistent clustering schemes. In the context of the standard statistical learning theory, the three conditions in the theorem are rather natural. The first condition mainly takes care of the estimation error. Implicitly, it restricts the size of the function class $\mathcal{F}_n$ by incorporating the shattering coefficient. We decided to state condition 1 in this rather abstract way to make the theorem as general as possible. We will see later how it can be used in concrete applications. Of course, there are many more ways to specify the size of function classes, and many of them might lead to better bounds in the end. However, in this paper we are not so much concerned with obtaining the sharpest bounds, but we want to demonstrate the general concept (as the reader can see in appendix, the proofs are already long enough using simple shattering numbers). The second condition in the theorem takes care of the approximation error. Intuitively it is clear that if we want to approximate solutions in $\mathcal{F}$, eventually $\mathcal{F}_n$ needs to be "close" to $\mathcal{F}$. The third condition establishes a relation between the quality function Q and the distance function d: if two clusterings f and g are close with respect to d, then their quality values $Q(f)$ and $Q(g)$ are close, too. We need this property to be able to conclude from "closeness" as in Condition 2 to "closeness" of the clustering quality values.

Finally, we would like to point out a few technical treats. First of all, note that the function class $\mathcal{F}_n$ is allowed to be data dependent. Secondly, as opposed to most results in empirical risk minimization we do not assume that Q_n is an unbiased estimator of Q (that is, we allow $\mathbb{E}Q_n \neq Q$), nor does Q need to be "an expectation" (that is, of the form $Q(f) = \mathbb{E}(\Omega(f,X))$ for some Ω). Both facts make the proof more technical, as many of the standard tools (symmetrization, concentration inequalities) become harder to apply. However, this is necessary since in the context of clustering biased estimators pop up all over the place. We will see that many of the popular clustering objective functions lead to biased estimators.

3. Nearest Neighbor Clustering—General Theory

The theorem presented in the last section shows sufficient conditions under which clustering can be performed consistently. Now we want to present a generic algorithm which can be used to minimize arbitrary clustering objective functions. With help of Theorem 1 we can then prove the consistency of its results for a large variety of clustering objective functions.

We have seen that the key to obtain consistent clustering schemes is to work with an appropriate function class. But of course, given quality functions Q and Q_n, the question is how such a function space can be constructed in practice. Essentially, three requirements have to be satisfied:

- The function space $\mathcal{F}_n$ has to be "small". Ideally, it should only contain polynomially many functions.

- The function space $\mathcal{F}_n$ should be "rich enough". In the limit $n \to \infty$, we would like to be able to approximate any (reasonable) measurable function.
- We need to be able to solve the optimization problem $\operatorname{argmin}_{f \in \mathcal{F}_n} Q_n(f)$. This sounds trivial at first glance, but in practice is far from easy.

One rather straightforward way to achieve all requirements is to use a function space of piecewise constant functions. Given a partitioning of the data space in small cells, we only look at clusterings which are constant on each cell (that is, the clustering never splits a cell). If we make sure that the number of cells is only of the order $\log(n)$, then we know that the number of clusterings is at most $K^{\log(n)} = n^{\log(K)}$, which is polynomial in n. In the following we will introduce a data-dependent random partition of the space which turns out to be very convenient.

3.1 Nearest Neighbor Clustering—The Algorithm

We will construct a function class $\mathcal{F}_n$ as follows. Given a finite sample $X_1, \ldots, X_n \in \mathbb{R}^d$, the number K of clusters to construct, and a number $m \in \mathbb{N}$ with $K \leq m \ll n$, randomly pick a subset of m "seed points" $X_{s_1}, \ldots, X_{s_m}$. Assign all other data points to their closest seed points, that is for all $j = 1, \ldots, m$ define the set Z_j as the subset of data points whose nearest seed point is X_{s_j}. In other words, the sets $Z_1, \ldots, Z_m$ are the Voronoi cells induced by the seeds $X_{s_1}, \ldots, X_{s_m}$. Then consider all partitions of $\mathcal{X}_n$ which are constant on all the sets $Z_1, ..., Z_m$. More formally, for given seeds we define the set $\mathcal{F}_n$ as the set of all functions

$$\mathcal{F}_n := \{ f : \mathcal{X} \to \{1, \ldots, K\} \mid \forall\, j = 1, \ldots, m : \forall z, z' \in Z_j : f(z) = f(z') \}.$$

Obviously, the function class $\mathcal{F}_n$ contains K^m functions, which is polynomial in n if the number m of seeds satisfies $m \in O(\log n)$. Given $\mathcal{F}_n$, the most simple polynomial-time optimization algorithm is then to evaluate $Q_n(f)$ for all $f \in \mathcal{F}_n$ and choose the solution $f_n = \operatorname{argmin}_{f \in \mathcal{F}_n} Q_n(f)$. We call the resulting clustering the *nearest neighbor clustering* and denote it by $\mathrm{NNC}(Q_n)$. The entire algorithm is summarized in Figure 1. We have already published results on the empirical performance

Nearest Neighbor Clustering NNC(Q_n), naive implementation

Parameters: number K of clusters to construct, number $m \in \mathbb{N}$ of seed points to use (with $K \leq m \ll n$), clustering quality function Q_n

Input: data set $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, distances $d_{ij} = d(X_i, X_j)$

- Subsample m seed points from the data points, without replacement.
- Build the Voronoi decomposition $Z_1, \ldots, Z_m$ of $\mathcal{X}_n$ based on the distances d_{ij} using the seed points as centers
- Define $\mathcal{F}_n := \{ f : \mathcal{X}_n \to \{1, \ldots, K\} \mid f \text{ constant on all cells } Z_j \}$
- For all $f \in \mathcal{F}_n$ evaluate $Q_n(f)$.

Output: $f_n := \operatorname{argmin}_{f \in \mathcal{F}_n} Q_n(f)$

Figure 1: Nearest neighbor clustering for a general clustering objective function Q_n.

of the algorithm in von Luxburg et al. (2008), and more results can be found in Section 3 of Jegelka (2007). We have found that on finite samples, the algorithm performs surprisingly well in terms of quality function: using $m = \log n$ seed points, the objective function values obtained at the solutions are comparable to these of K-means or spectral clustering, respectively. Moreover, there exist efficient ways to compute f_n using branch and bound methods. Using these methods, the running time of nearest neighbor clustering using $m = \log n$ seeds is roughly comparable to the one of the other clustering algorithms. See von Luxburg et al. (2008) and Jegelka (2007) for details on the experimental results.

3.2 Consistency of Nearest Neighbor Clustering (General Statement)

Now we want to prove that nearest neighbor clustering is consistent. We will see that even though we can rely on Theorem 1, the consistency proof for nearest neighbor clustering does not come for free. Let $f : \mathcal{X} \to \{1,\dots,K\}$ be a clustering function. In the following, we will often use the notation f_k for the indicator function of the k-th cluster:

$$f_k(x) := \mathbb{1}_{f(x)=k}.$$

This is a slight abuse of notation, as we already reserved the notation f_n for the minimizer of the empirical quality function. However, from the context it will always be clear whether we will refer to f_n or f_k, respectively, as we will not mix up the letters n (for the sample size) and k (a cluster index).

As distance function between two clusterings we use the 0-1-loss

$$d(f,g) := \mathbb{P}(f(X) \neq g(X) | X_1,\dots,X_n).$$

Here the conditioning is needed for the cases where the functions f or g are data dependent. Note that in clustering, people often consider a variant of this distance which is independent with respect to the choice of labels, that is they choose $\tilde{d}(f,g) := \min_\pi P(f(X) \neq \pi(g(X)) | X_1,\dots,X_n)$, where π runs over all permutations of the set $\{1,\dots,K\}$. However, we will see that for our purposes it does not hurt to use the overly sensitive 0-1 distance instead. The main reason is that at the end of the day, we only want to compare functions based on their quality values, which do not change under label permutations. In general, the theorems and proofs could also be written in terms of $\tilde{d}$. For better readability, we decided to stick to the standard 0-1 distance, though.

We will see below that in many cases, even in the limit case one would like to use a function space $\mathcal{F}$ which is a proper subset of $\mathcal{H}$. For example, one could only be interested in clusterings where all clusters have a certain minimal size, or where the functions satisfy certain regularity constraints. In order to be able to deal with such general function spaces, we will introduce a tool to restrict function classes to functions satisfying certain conditions. To this end, let

$$\Phi : \mathcal{H} \to \mathbb{R}^+$$

be a functional which quantifies certain aspects of a clustering. In most cases, we will use functionals Φ which operate on the individual cluster indicator functions f_k. For example, $\Phi(f_k)$ could measure the size of cluster k, or the smoothness of the cluster boundary. The function class $\mathcal{F}$ will then be defined as

$$\mathcal{F} = \{ f \in \mathcal{H} \mid \Phi(f_k) > a \text{ for all } k = 1,\dots,K \},$$

where $a \geq 0$ is a constant. In general, the functional Φ can be used to encode our intuition about "what a cluster is". Note that this setup also includes the general case of $\mathcal{F} = \mathcal{H}$, that is the case where we do not want to make any further restrictions on $\mathcal{F}$, for example by setting $\Phi(f_k) \equiv 1$, $a \equiv 0$. As it is the case for Q, we will usually not be able to compute Φ on a finite sample only. Hence we also introduce an empirical counterpart Φ_n which will be used in the finite sample case.

The following theorem will state sufficient conditions for the consistency of nearest neighbor clustering. For simplicity we state the theorem for the case $\mathcal{X} = \mathbb{R}^d$, but the proofs can also be carried over to more general spaces. Also, note that we only state the theorem for the case $d \geq 2$; in case $d = 1$ the theorem holds as well, but the formulas look a bit different.

Theorem 2 (Consistency of nearest neighbor clustering) *Let $\mathcal{X} = \mathbb{R}^d$, $d \geq 2$, $Q : \mathcal{H} \to \mathbb{R}^+$ be a clustering quality function, and $Q_n : \mathcal{H} \to \mathbb{R}^+$ an estimator of this function which can be computed based on the finite sample only. Similarly, let $\Phi : \mathcal{H} \to \mathbb{R}^+$, and $\Phi_n : \mathcal{H} \to \mathbb{R}^+$ an estimator of this function. Let $a > 0$ and $(a_n)_{n\in\mathbb{N}}$ be such that $a_n > a$ and $a_n \to a$. Let $m = m(n) \leq n \in \mathbb{N}$. Finally, denote $d(f,g)$ the 0-1-loss, and let $NN_m(x)$ be the nearest neighbor of x among $X_1, \ldots, X_m$ according to the Euclidean distance. Define the function spaces*

$$\begin{aligned}
\mathcal{F} &:= \{ f : \mathbb{R}^d \to \{1,\ldots,K\} \mid f \text{ continuous a.e. and } \forall k \in \{1,\ldots,K\}\ \ \Phi(f_k) > a \} \\
\mathcal{F}_n &:= \{ f : \mathbb{R}^d \to \{1,\ldots,K\} \mid f(x) = f(NN_m(x)) \text{ and } \forall k \in \{1,\ldots,K\}\ \ \Phi_n(f_k) > a_n \} \\
\widetilde{\mathcal{F}_n} &:= \bigcup_{X_1,\ldots,X_n \in \mathbb{R}^d} \mathcal{F}_n \\
\widehat{\mathcal{F}_n} &:= \{ f : \mathbb{R}^d \to \{1,\ldots,K\} \mid \exists \text{ Voronoi partition of } m \text{ cells: } f \text{ constant on all cells} \}.
\end{aligned}$$

Assume that the following conditions are satisfied:

1. *$Q_n(f)$ is a consistent estimator of $Q(f)$ which converges sufficiently fast for all $f \in \widetilde{\mathcal{F}_n}$:*
$$\forall \varepsilon > 0, K^m (2n)^{(d+1)m^2} \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| > \varepsilon) \to 0,$$

2. *$\Phi_n(f_k)$ is a consistent estimator of $\Phi(f_k)$ which converges sufficiently fast for all $f \in \widehat{\mathcal{F}_n}$:*
$$\forall \varepsilon > 0, K^m (2n)^{(d+1)m^2} \sup_{f \in \widehat{\mathcal{F}_n}} \mathbb{P}(|\Phi_n(f_k) - \Phi(f_k)| > \varepsilon) \to 0,$$

3. *Q is uniformly continuous with respect to the pseudo-distance $d(f,g)$ between $\mathcal{F}$ and $\widetilde{\mathcal{F}_n}$, as defined in Condition (3) of Theorem 1,*

4. *$\Phi_k(f) := \Phi(f_k)$ is uniformly continuous with respect to the pseudo-distance $d(f,g)$ between $\mathcal{F}$ and $\widehat{\mathcal{F}_n}$, as defined in Condition (3) of Theorem 1,*

5. *a_n decreases slowly enough to a:*
$$K^m (2n)^{(d+1)m^2} \sup_{g \in \widehat{\mathcal{F}_n}, k} \mathbb{P}(\Phi_n(g_k) - \Phi(g_k) \geq a_n - a) \to 0,$$

6. $m \to \infty$.

Then nearest neighbor clustering based on m seed points using quality function Q_n is weakly consistent, that is for $f_n \in \operatorname{argmin}_{f \in \mathcal{F}_n} Q_n(f)$ and $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} Q(f)$ we have $Q(f_n) \to Q(f^*)$ in probability.

This theorem is still rather abstract, but pretty powerful. In the following we will demonstrate this by applying it to many concrete clustering objective functions. To define our objective functions, we will from now on adopt the convention $0/0 = 0$.

4. Nearest Neighbor Clustering with Popular Clustering Objective Functions

In this section we want to study the consistency of nearest neighbor clustering when applied to particular objective functions. For simplicity we assume in this section that $\mathcal{X} = \mathbb{R}^d$.

4.1 NNC Using the K-means Objective Function

The K-means objective function is the within-cluster sum of squared distances, called WSS for short. To define it properly, for a given clustering function $f : \mathbb{R}^d \to \{1, \ldots, K\}$ we introduce the following quantities:

$$\mathrm{WSS}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} f_k(X_i) \|X_i - c_{k,n}\|^2 \qquad \text{where}$$

$$c_{k,n} := \frac{1}{n_k} \frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i \qquad \text{and} \qquad n_k := \frac{1}{n} \sum_{i=1}^{k} f_k(X_i)$$

$$\mathrm{WSS}(f) := \mathbb{E} \sum_{k=1}^{K} f_k(X) \|X - c_k\|^2 \qquad \text{where} \qquad c_k := \frac{\mathbb{E} f_k(X) X}{\mathbb{E} f_k(X)}.$$

Here, WSS_n plays the role of Q_n and WSS the role of Q. Let us point out some important facts. First the empirical quality function is not an unbiased estimator of the true one, that is $\mathbb{E}\mathrm{WSS}_n \neq \mathrm{WSS}$ and $\mathbb{E} c_{k,n} \neq c_k$ (note that in the standard treatment of K-means this can be achieved, but not on arbitrary function classes, see below for some discussion). However, at least we have $\mathbb{E} n_k = \mathbb{E} f_k(X)$ and $\mathbb{E} \frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i = \mathbb{E} f_k(X) X$. Moreover, one should remark that if we define $\mathrm{WSS}(\cdot, \mathbb{P}) := \mathrm{WSS}$ then $\mathrm{WSS}_n = \mathrm{WSS}(\cdot, \mathbb{P}_n)$ where $\mathbb{P}_n$ is the empirical distribution.

Secondly, our setup for proving the consistency of nearest neighbor clustering with the WSS objective function is considerably more complicated than proving the consistency of the global minimizer of the K-means algorithm (e.g., Pollard, 1981). The reason is that for the K-means algorithm one can use a very helpful equivalence which does not hold for nearest neighbor clustering. Namely, if one considers the minimizer of WSS_n in the space of *all possible partitions*, then one can see that the clustering constructed by this minimizer always builds a Voronoi partition with K cells; the same holds in the limit case. In particular, given the cluster centers $c_{k,n}$ one can reconstruct the whole clustering by assigning each data point to the closest cluster center. As a consequence, to prove the convergence of K-means algorithms one usually studies the convergence of the empirical cluster centers $c_{k,n}$ to the true centers c_k. However, in our case this whole chain of arguments breaks

down. The reason is that the clusters chosen by nearest neighbor clustering *from the set* $\mathcal{F}_n$ are not necessarily Voronoi cells, they do not even need to be convex (all clusters are composed by small Voronoi cells, but the union of "small" Voronoi cells is not a "large" Voronoi cell). Also, it is not the case that each data point is assigned to the cluster corresponding to the closest cluster center. It may very well happen that a point x belongs to cluster C_i, but is closer to the center of another cluster C_j than to the center of its own cluster C_i. Consequently, we cannot reconstruct the nearest neighbor clustering from the centers of the clusters. This means that we cannot go over to the convergence of centers, which makes our proof considerably more involved than the one of the standard K-means case.

Due to these technical problems, it will be of advantage to only consider clusters which have a certain minimal size (otherwise, the cluster quality function WSS is not uniformly continuous). To achieve this, we use the functionals

$$\Phi_{\mathrm{WSS}}(f_k) := \mathbb{E} f_k(X), \qquad \Phi_{\mathrm{WSS}_n}(f_k) := n_k(f).$$

and will only consider clusterings where $\Phi(f_k) \geq a > 0$. In practice, this can be interpreted as a simple means to avoid empty clusters. The constant a can be chosen so small that its only effect is to make sure that each cluster contains at least one data point. The corresponding function spaces are

$$\mathcal{F} := \{f : \mathbb{R}^d \to \{1,\dots,K\} \mid f \text{ continuous a.e. and } \forall k \in \{1,\dots,K\} \ \Phi_{\mathrm{WSS}}(f_k) > a\}$$
$$\mathcal{F}_n := \{f : \mathbb{R}^d \to \{1,\dots,K\} \mid f(x) = f(NN_m(x)) \text{ and } \forall k \in \{1,\dots,K\} \ \Phi_{\mathrm{WSS}_n}(f_k) > a_n\}$$

Moreover, for technical convenience we restrict our attention to probability measures which have a bounded support inside some large ball, that is which satisfy $\operatorname{supp} \mathbb{P} \subset B(0,A)$ for some constant $A > 0$. It is likely that our results also hold in the general case, but the proof would get even more complicated. With the notation of Theorem 2 we have:

Theorem 3 (Consistency of NNC(WSS)**)** *Assume that* $a_n > a, a_n \to a, m \to \infty$ *and*

$$\frac{m^2 \log n}{n(a-a_n)^2} \to 0.$$

Then for all probability measures on $\mathbb{R}^d$ *with bounded support, nearest neighbor clustering with* WSS *is consistent, that is if* $n \to \infty$ *then* $\mathrm{WSS}(f_n) \to \mathrm{WSS}(f^*)$ *in probability.*

This theorem looks very nice and simple. The conditions on a_n and m are easily satisfied as soon as these quantities do not converge too fast. For example, if we define

$$a_n = a + \frac{1}{\log n} \quad \text{and} \quad m = \log n$$

then

$$\frac{m^2 \log n}{n(a_n-a)^2} = \frac{(\log n)^5}{n} \to 0.$$

Moreover, it is straightforward to see from the proofs that this theorem is still valid if we consider the objective functions WSS_n and WSS with $\|\cdot\|$ instead of $\|\cdot\|^2$. It also holds for any other norm, such as the p-norms $\|\cdot\|_p$. However, it does not necessarily hold for powers of norms (in this sense, the squared Euclidean norm is an exception). The proof shows that the most crucial property is

$$\|X_i - c_{k,n}\| - \|X_i - c_k\| \leq \text{const} \cdot \|c_{k,n} - c_k\|.$$

This is straightforward if the triangle inequality holds, but might not be possible for general powers of norms.

By looking more carefully at our proofs one can state the following rate of convergence:

Theorem 4 (Convergence Rate for NNC(WSS)**)** *Assume that* $\operatorname{supp}\mathbb{P} \subset B(0,A)$ *for some constant* $A > 0$ *and that* $n(a_n - a)^2 \to \infty$. *Let* $\varepsilon \leq 1$ *and* $a^* := \inf_k \mathbb{E} f_k^*(X) - a > 0$. *Then there exists :*

$$\begin{aligned} &N = N((a_n), a^*) \in \mathbb{N}, \\ &C_1 = C_1(a, a^*, \varepsilon, K, A) > 0, && C_2 = C_2(a, a^*, \varepsilon, A, f^*, \mathbb{P}) > 0, \\ &C_3 = C_3(a, d, \varepsilon, K, A) > 0, && C_4 = C_4(a, d, A) > 0 \end{aligned}$$

such that for $n \geq N$ *the following holds true:*

$$\begin{aligned} &\mathbb{P}(|\,\mathrm{WSS}(f_n) - \mathrm{WSS}(f^*)| \geq \varepsilon) \\ &\leq C_1 e^{-C_2 m} + K^{m+1}(2n)^{(d+1)m^2} \left(C_3 e^{-C_4 \varepsilon^2 n} + 8K e^{-\frac{n(a_n-a)^2}{8}} \right). \end{aligned}$$

At first glance, it seems very tempting to try to use the Borel-Cantelli lemma to transform the weak consistency into strong consistency. However, we do not have an explicit functional form of dependency of C_2 on ε. The main reason is that in Lemma 11 (Appendix) the constant $b(\varepsilon)$ will be defined only implicitly. If one would like to prove strong consistency of nearest neighbor clustering with WSS one would have to get an explicit form of $b(\varepsilon)$ in Lemma 11.

For a general discussion relating the consistency result of NNC(WSS) in to the consistency results by Pollard (1981) and others see Section 5.

4.2 NNC Using Standard Graph-cut Based Objective Functions

In this section we want to look into the consistency of nearest neighbor clustering for graph based objective functions as they are used in spectral clustering (see von Luxburg, 2007 for details). Let $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be a similarity function which is upper bounded by a constant C. The two main quantities we need to define graph-cut based objective functions are the cut and the volume. For a given cluster described by the cluster indicator function $f_k : \mathbb{R}^d \to \{0,1\}$, we set

$$\mathrm{cut}(f_k) := \mathrm{cut}(f_k, \mathbb{P}) := \mathbb{E} f_k(X_1)(1 - f_k(X_2))s(X_1, X_2),$$

$$\mathrm{vol}(f_k) := \mathrm{vol}(f_k, \mathbb{P}) := \mathbb{E} f_k(X_1)s(X_1, X_2).$$

For $f \in \mathcal{H}$ we can then define the normalized cut and the ratio cut by

$$\mathrm{Ncut}(f) := \mathrm{Ncut}(f, \mathbb{P}) := \sum_{k=1}^{K} \frac{\mathrm{cut}(f_k)}{\mathrm{vol}(f_k)},$$
$$\mathrm{RatioCut}(f) := \mathrm{RatioCut}(f, \mathbb{P}) := \sum_{k=1}^{K} \frac{\mathrm{cut}(f_k)}{\mathbb{E} f_k(X)}.$$

The empirical estimators of these objective functions will be $\mathrm{Ncut}(f, \mathbb{P}_n)$ and $\mathrm{RatioCut}(f, \mathbb{P}_n)$, in explicit formulas:

$$\mathrm{cut}_n(f_k) := \frac{1}{n(n-1)} \sum_{i,j=1}^{n} f_k(X_i)(1 - f(X_j))s(X_i, X_j),$$
$$\mathrm{vol}_n(f_k) := \frac{1}{n(n-1)} \sum_{i,j=1}^{n} f_k(X_i)s(X_i, X_j), \qquad n_k := \frac{1}{n} \sum_{i=1}^{k} f_k(X_i),$$
$$\mathrm{Ncut}_n(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}_n(f_k)}{\mathrm{vol}_n(f_k)}, \qquad \mathrm{RatioCut}_n(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}_n(f_k)}{n_k}.$$

Again we need to define how we will measure the size of the clusters. We will use

$$\Phi_{\mathrm{cut}}(f_k) := \mathrm{vol}(f_k), \qquad \Phi_{\mathrm{Ncut}}(f_k) := \mathrm{vol}(f_k), \qquad \Phi_{\mathrm{RatioCut}}(f_k) := \mathbb{E} f_k(X).$$

with the corresponding empirical quantities $\Phi_{\mathrm{cut}_n}, \Phi_{\mathrm{Ncut}_n}$ and $\Phi_{\mathrm{RatioCut}_n}$. Then, with the notations of Theorem 2, we have:

Theorem 5 (Consistency of $\mathrm{NNC}(\mathrm{cut}), \mathrm{NNC}(\mathrm{Ncut})$ **and** $\mathrm{NNC}(\mathrm{RatioCut})$**)** *Assume that the similarity function s is bounded by a constant $C > 0$, let $a_n > a$, $a_n \to a$, $m \to \infty$ and*

$$\frac{m^2 \log n}{n(a - a_n)^2} \to 0.$$

Then nearest neighbor clustering with $\mathrm{cut}, \mathrm{Ncut}$ *and* $\mathrm{RatioCut}$ *is universally weakly consistent, that is for all probability measures, if $n \to \infty$ we have* $\mathrm{cut}(f_n) \to \mathrm{cut}(f^*)$, $\mathrm{Ncut}(f_n) \to \mathrm{Ncut}(f^*)$ *and* $\mathrm{RatioCut}(f_n) \to \mathrm{RatioCut}(f^*)$ *in probability.*

For these objective functions one can also state a rate of convergence. For sake of shortness we only state it for the normalized cut:

Theorem 6 (Convergence Rate for $\mathrm{NNC}(\mathrm{Ncut})$**)** *Assume that the similarity function s is bounded by $C > 0$ and that $n(a_n - a)^2 \to \infty$. Let $\varepsilon \leq 1$ and $a^* := \inf_k vol(f_k^*) - a > 0$. Then there exist*

$$N = N((a_n), a^*) \in \mathbb{N},$$
$$C_1 = C_1(a, a^*, \varepsilon, K, C) > 0, \qquad C_2 = C_2(a, a^*, \varepsilon, C, K, f^*, \mathbb{P}) > 0,$$
$$C_3 = C_3(a, \varepsilon, K, C) > 0, \qquad C_4 = C_4(a, K, C) > 0.$$

such that for $n \geq N$ the following holds true:

$$\mathbb{P}(|\mathrm{Ncut}(f_n) - \mathrm{Ncut}(f^*)| \geq \varepsilon)$$
$$\leq C_1 e^{-C_2 m} + K^{m+1}(2n)^{(d+1)m^2} \left(C_3 e^{-C_4 \varepsilon^2 n} + 8Ke^{-\frac{n(a_n - a)^2}{8}} \right).$$

4.3 NNC Using the Modularity Objective Function

A slightly different objective functions for graph clustering is the "modularity", which has been put forward by Newman (2006) for detecting communities in networks. In this paper, the modularity is formulated as an objective function to find communities in a finite graph. However, as it is the case for Ncut or RatioCut, the modularity cannot be directly minimized. Instead, a spectral relaxation has been developed to minimize the modularity, see Newman (2006) for details. Of course, the nearest neighbor clustering algorithm can also be used to minimize this objective function directly, without using a relaxation step. Using our own notation we define:

$$\mathrm{Mod}_n(f) = \sum_{k=1}^{n} \frac{1}{n(n-1)} \sum_{i \neq j} f_k(X_i) f_k(X_j) \left(\frac{1}{(n-1)^2} \sum_{l,l \neq i} s(X_i, X_l) \sum_{l,l \neq j} s(X_j, X_l) - s(X_i, X_j) \right),$$

$$\mathrm{Mod}(f) = \sum_{k=1}^{n} \int \int f_k(X) f_k(Y) \left(\int s(X,Z) d\mathbb{P}(Z) \int s(Y,Z) d\mathbb{P}(Z) - s(X,Y) \right) d(\mathbb{P} \times \mathbb{P})(X,Y).$$

In the proof we will see that as the limit function $\mathrm{Mod}(\cdot)$ is uniformly continuous on $\mathcal{H}$, we do not need to quantify any function Φ or Φ_n to measure the volume of the clusters. The function classes are thus

$$\mathcal{F} := \{ f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f \text{ continuous a.e.} \},$$
$$\mathcal{F}_n := \{ f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f(x) = f(NN_m(x)) \}.$$

Theorem 7 (Consistency of NNC(Mod)**)** *Assume that $m \to \infty$ and*

$$\frac{m^2 \log n}{n} \to 0.$$

Then nearest neighbor clustering with Mod *is universally weakly consistent: for all probability measures, if $n \to \infty$ then $\mathrm{Mod}(f_n) \to \mathrm{Mod}(f^*)$ in probability.*

4.4 NNC Using Objective Function Based on the Ratio of Within-cluster and Between-cluster Similarity

Often, clustering algorithms try to minimize joint functions of the within-cluster similarity and the between cluster similarity. The most popular choice is the ratio of those two quantities, which is closely related to the criterion used in Fisher linear discriminant analysis. Formally, the between-cluster similarity corresponds to the cut, and the within similarity of cluster k is given by

$$\mathrm{WS} := \mathbb{E} f(X_1) f(X_2) s(X_1, X_2).$$

Thus the ratio of between- and within-cluster similarity is given as

$$\mathrm{BWR}(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}(f_k)}{\mathrm{WS}(f_k)}.$$

Again we use their empirical estimations:

$$\mathrm{WS}_n(f_k) := \frac{1}{n(n-1)} \sum_{i,j=1}^{n} f_k(X_i) f_k(X_j) s(X_i, X_j),$$
$$\mathrm{BWR}_n(f) := \sum_{k=1}^{K} \frac{\mathrm{cut}_n(f_k)}{\mathrm{WS}_n(f_k)}.$$

To measure the size of the cluster we use

$$\Phi_{\mathrm{BWR}}(f_k) := \mathrm{WS}(f_k)$$

and its natural empirical counterpart. This leads to function spaces

$$\mathcal{F} := \{f : \mathbb{R}^d \to \{1,\dots,K\} \mid f \text{ continuous a.e. and } \forall k \in \{1,\dots,K\} \ \Phi_{\mathrm{BWR}}(f_k) > a\},$$
$$\mathcal{F}_n := \{f : \mathbb{R}^d \to \{1,\dots,K\} \mid f(x) = f(NN_m(x)) \text{ and } \forall k \in \{1,\dots,K\} \ \Phi_{\mathrm{BWR}_n}(f_k) > a_n\}.$$

Theorem 8 (Consistency of NNC(BWR)**)** *Assume that the similarity function s is bounded by a constant $C > 0$, let $a_n > a, a_n \to a, m \to \infty$ and*

$$\frac{m^2 \log n}{n(a-a_n)^2} \to 0.$$

Then nearest neighbor clustering with BWR *is universally weakly consistent, that is for all probability measure if $n \to \infty$ then* $\mathrm{BWR}(f_n) \to \mathrm{BWR}(f^*)$ *in probability.*

5. Relation to Previous Work

In this section we want to discuss our results in the light of the existing literature on consistent clusterings.

5.1 Standard Consistency Results for Center-based Algorithms

For a few clustering algorithms, consistency results are already known. The most well-known among them is the K-means algorithm. For this algorithm it has been first proved by Pollard (1981) that the global minimizer of the K-means objective function on a finite sample converges to the global minimizer on the underlying space.

First of all, we would like to point out that the consistency result by Pollard (1981) can easily be recovered using our theorems. Let us briefly recall the standard K-means setting. The objective function which K-means attempts to optimize is the function WSS, which we already encountered in the last sections. In the standard K-means setting the optimization problem is stated over the space of all measurable functions $\mathcal{H}$:

$$f^* = \operatorname*{argmin}_{f \in \mathcal{H}} \mathrm{WSS}(f).$$

It is not difficult to prove that the solution f^* of this optimization problem always has a particular form. Namely, the solution f^* forms a Voronoi decomposition of the space, where the cluster

centers c_k are the centers of the Voronoi cells. Thus, we can rewrite the optimization problem above equivalently as

$$f^* = \operatorname*{argmin}_{f \in \mathcal{G}_K} \mathrm{WSS}(f)$$

where $\mathcal{G}_K$ denotes the set of all clusterings for which the clusters are Voronoi cells. The optimization problem for the finite sample case can be stated analogously:

$$f_n = \operatorname*{argmin}_{f \in \mathcal{G}_K} \mathrm{WSS}_n(f).$$

So in this particular case we can set $\mathcal{F}_n = \mathcal{F} = \mathcal{G}_K$. This means that even though the original optimization problem has been set up to optimize over the huge set $\mathcal{H}$, the optimization only needs to run over the small set $\mathcal{G}_K$. It is well known that the shattering coefficient of $\mathcal{G}_K$ is polynomial in n, namely it is bounded by $K^K n^{(d+1)K^2}$ (cf. Lemma 10). Moreover, the uniform continuity of WSS on $\mathcal{G}_K$ (Condition (3) of Theorem 2) can easily be verified if we assume that the probability distribution has compact support. As a consequence, using similar techniques as in the proofs of Theorem 3 we can prove that the global minimizer of the empirical K-means objective function WSS_n converges to the global minimizer of the true K-means objective function WSS. By this we recover the well-known result by Pollard (1981), under slightly different assumptions. In this sense, our Theorem 1 can be seen as a blueprint for obtaining Pollard-like results for more general objective functions and function spaces.

Are there any more advantages of Theorem 3 in the K-means setting? At first glance, our result in Theorem 3 looks similar to Pollard's result: the global minimizers of both objective functions converge to the true global minimizer. However, in practice there is one important difference. Note that as opposed to many vector quantization problems (cf. Garey et al., 1982), minimizing the K-means objective function is not NP-hard in n: the solution is always a Voronoi partition, there exist polynomially many Voronoi partitions of n points, and they can be enumerated in polynomial time (cf. Inaba et al., 1994). However, the size of the function class $\mathcal{G}_K$ is still so large that it would take too long to simply enumerate all its functions and select the best one. Namely, we will see in Lemma 10 that the number of Voronoi partitions of n points in $\mathbb{R}^d$ using K cells is bounded by $n^{(d+1)K}$, which is huge even for moderate d and K. As a work-around in practice one uses the well-known K-means *algorithm*, which is only able to find a *local* minimum of $\mathrm{WSS}_n(f)$. In contrast, nearest neighbor clustering works with a different function class which is much smaller than $\mathcal{G}_K$: it has only size $n^{\log K}$. On this smaller class we are still able to compute the *global* minimum of $\mathrm{WSS}_n(f)$. Consequently, our result in Theorem 3 is not only a theoretical statement about some abstract quantity as it is the case for Pollard's result, but it applies to the algorithm used in practice. While Pollard's result abstractly states that the global minimum (which cannot be computed efficiently) converges, our result implies that the result of nearest neighbor clustering does converge.

5.2 Consistency of Spectral Clustering

In the previous section we have seen in Theorems 5 and 6 that NNC is consistent for all the standard graph cut objective functions. Now we want to discuss these results in connection with the graph cut literature. It is well known that the discrete optimization problem of minimizing Ncut_n or $\mathrm{RatioCut}_n$ is an NP-hard problem, see Wagner and Wagner (1993). However, approximate solutions of relaxed

problems can be obtained by spectral clustering, see von Luxburg (2007) for a tutorial. Consistency results for spectral clustering algorithms have been proved in von Luxburg et al. (2008). These results show that under certain conditions, the solutions computed by spectral clustering on finite samples converge to some kind of "limit solutions" based on the underlying distribution. In the light of the previous discussions, this sounds plausible, as the space of solutions of spectral clustering is rather restricted: we only allow solutions which are eigenfunctions of certain integral operators. Thus, spectral clustering implicitly works with a small function class.

However, it is important to note that the convergence results of spectral clustering do not make any statement about the minimizers of Ncut (a similar discussion also holds for RatioCut). The problem is that on any finite sample, spectral clustering only solves a relaxation of the original problem of minimizing Ncut_n. The Ncut_n-value of this solution can be arbitrarily far away from the minimal Ncut_n-value on this sample (Guattery and Miller, 1998), unless one makes certain assumptions which are not necessarily satisfied in a standard statistical setting (cf. Spielman and Teng, 1996, or Kannan et al., 2004). Thus the convergence statements for the results computed by the spectral clustering algorithm cannot be carried over to consistency results for the minimizers of Ncut. One knows that spectral clustering converges, but one does not have any guarantee about the Ncut-value of the solution. Here our results for nearest neighbor clustering present an improvement, as they directly refer to the minimizer of Ncut. While it is known that spectral clustering converges to "something", for the solutions computed by nearest neighbor clustering we know that they converge to the global minimizer of Ncut (or RatioCut, respectively).

5.3 Consistency of Other Clustering Schemes

To the best of our knowledge, apart from results on center-based algorithms and spectral clustering, there are very few non-parametric clustering algorithms for which statistical consistency has been proved so far. The only other major class of algorithms for which consistency has been investigated is the class of linkage algorithms. While single linkage can be proved to be "fractionally consistent", that is it can at least discover sufficiently distinct high-density regions, both complete and average linkage are not consistent and can be misleading (cf. Hartigan, 1981, 1985). A more general method for hierarchical clustering used in Wong and Lane (1983) is statistically consistent, but essentially first estimates the density and then constructs density level sets based on this estimator.

Concerning parametric clustering algorithms, the standard setting is a model-based approach. One assumes that the underlying probability distribution has a certain parametric form (for example a mixture of Gaussians), and the goal is to estimate the parameters of the distribution from the sample. Estimating parameters in parametric models has been intensively investigated in statistics, in particular in the maximum likelihood framework and the Bayesian framework (for an overview how this can be done for clustering see Fraley and Raftery, 1998, or the book McLachlan and Peel, 2004). Numerous consistency results are known, but typically they require that the true underlying distribution indeed comes from the model class under consideration. For example, in a Bayesian setting one can show that in the large sample limit, the posterior distribution will concentrate around the true mixture parameters. However, if the model assumptions are not satisfied, counter-examples to consistency can be constructed. Moreover, the consistency results mentioned above are theoretic in the sense that the algorithm used in practice does not necessarily achieve them. Standard approaches to estimate mixture parameters are the EM algorithm (in a frequentist of MAP setting), or for example Markov Chain Monte Carlo sampling in a fully Bayesian approach. However, as it is the case for

the K-means algorithm, these methods can get stuck in local optima, and no convergence towards the global optimum can be guaranteed. Another way to tackle model-based clustering problems is based on the minimum message length or minimum description length principle. The standard reference for MML approaches to learn mixtures is Figueiredo and Jain (2002), for a general overview on MDL see Grünwald (2007). Consistency results for MML are quite similar to the ones for the Bayesian approach: if the true distribution indeed comes from the mixture class and the number of components is known, then consistency can be achieved. For general results on consistency of MDL see Sections 16 and 17.11 in Grünwald (2007). Often, MML/MDL approaches are interpreted as a particular way to work with small function classes, consisting of functions which can be described in a "compact" way. In this sense, this method can also be seen as a way of achieving "small" function classes.

5.4 Sublinear Time Algorithms Using Subsampling

Some algorithms related to our approach have been published in the theoretical computer science community, such as Indyk (1999), Mishra et al. (2001), or Czumaj and Sohler (2007). The general idea is to use subsampling approaches to approximate clustering solutions, and to prove that these approximations are quite accurate. Given a sample of n points, one draws a subsample of $m \ll n$ points, applies some (approximate) clustering algorithm to the subsample, and then extends this clustering to the remaining points. Using techniques such as concentration inequalities, Chernoff bounds or Hoeffding bounds, one can then prove that the resulting clustering approximates the best clustering on the original point set.

While at first glance, this approach sounds very similar to our nearest neighbor clustering, note that the focus in these papers is quite a different one than ours. The authors do not aim for consistent clustering solutions (that is, solutions which are close to the "true clustering solution" of the underlying space), but they want to find algorithms to approximate the optimal clustering on a given finite sample in sublinear time. The sublinearity is achieved by the fact that already a very small subsample (say, $m = \log n$) is enough to achieve good approximation guarantees. However, our main point that it is important to control the size of the underlying function class, is not revealed in these papers. As the authors mainly deal with K-means type settings, they automatically work with polynomial function classes of center-based clusterings, and the issue of inconsistency does not arise. Moreover, subsampling is just one way of reducing the function class to a smaller size, there can be many others. In this sense, we believe that our "small function class" approach is more general than the subsampling approach.

Finally, one difference between our approach and the subsampling approach is the kind of results of interest. We are mainly concerned with asymptotic results, and on our way achieve approximation guarantees which are good for large sample size n. The focus of the subsampling papers is non-asymptotic, dealing with a small or moderate sample size n, and to prove approximation guarantees in this regime.

5.5 Other Statistical Learning Theory Approaches to Clustering

In the last years there have been several papers which started to look at clustering from a statistical learning theory perspective. A general statistical learning theory approach to clustering, based on a very similar intuition as ours, has already been presented in Buhmann (1998). Here the authors put forward an "empirical risk approximation" approach for unsupervised learning, along the lines

of empirical risk minimization for the supervised case. The setting under consideration is that the clustering quality function is an expectation with respect to the true underlying probability distribution, and the empirical quality function is the corresponding empirical expectation. Then, similar to the statistical learning theory for supervised learning, generalization bounds can be derived, for example using VC dimensions. Additionally, the authors discuss regularization approaches and relate them to annealing schemes for center-based clusterings.

A different approach has been investigated in Ben-David (2007). Here the author formalizes the notion of a "cluster description scheme". Intuitively, a clustering problem can be described by a cluster description scheme of size $l \in \mathbb{N}$ if each clustering can be described using l points from the space (and perhaps some additional parameter). For instance, this is the case for center-based clusterings, where the clustering can be described by the centroids only. Ben-David then proves generalization bounds for clustering description schemes which show that the global minimizer of the empirical quality function converges to the global minimizer of the true quality function. The proof techniques used in this paper are very close to the ones used in standard minimum description length results.

Another class of results about K-means algorithms has been proved in Rakhlin and Caponnetto (2007). After computing covering numbers for the underlying classes, the authors study the stability behavior of K-means. This leads to statements about the set of "almost-minimizers" (that is the set of all functions whose quality is ε close to the one of the global optimal solutions). As opposed to our results and all the other results discussed above, the main feature of this approach is that at the end of the day, one is able to make statements about the clustering functions themselves, rather than only about their quality values. In this sense, the approach in Rakhlin and Caponnetto (2007) has more powerful results, but its application is restricted to K-means type algorithms.

All approaches outlined above implicitly or explicitly rely on the same intuition as our approach: the function class needs to be "small" in order to lead to consistent clusterings. However, all previous results have some restrictions we could overcome in our approach. First of all, in the papers discussed above the quality function needs to be an expectation, and the empirical quality function is simply the empirical expectation. Here our results are more general: we neither require the quality functions to be expectations (for example, Ncut cannot be expressed as an expectation, it is a ratio of two expectations) nor do we require unbiasedness of the empirical quality function. Second, the papers discussed above make statements about global optimizers, but do not really deal with the question how such a global optimizer can be computed. The case of standard K-means shows that this is by no means simple, and in practice one has to use heuristics which discover local optima only. In contrast, we suggest a concrete algorithm (NNC) which computes the global optimum over the current function class, and hence our results not only concern abstract global minimizers which are hard to obtain, but refer to exactly the quantities which are computed by the algorithm. Finally, our algorithm has the advantage that it provides a framework for dealing with more general clustering objective functions than just center-based ones. This is not the case in the papers above.

Finally, we would like to mention that a rather general but vague discussion of some of the open issues in statistical approaches to clustering has been led in von Luxburg and Ben-David (2005). Our current paper partly solves some of the open issues raised there.

6. Discussion

Our paper is concerned with clustering algorithms which minimize certain quality functions. Our main point is that as soon as we require statistical consistency we have to work with function classes $\mathcal{F}_n$ which are "small". Our results have a similar taste as the well-known corresponding results for supervised classification. While in the domain of supervised classification practitioners are well aware of the effect of overfitting, it seems like this effect has been completely overlooked in the clustering domain.

We would like to highlight a convenient side-effect of working with small function classes. In clustering, for many objective functions the problem of finding the best partition of the discrete data set is an NP-hard problem (for example, this is the case for all balanced graph-cut objective functions). On the other side, if we restrict the function class $\mathcal{F}_n$ to have polynomial size (in n), then the trivial algorithm of evaluating all functions in $\mathcal{F}_n$ and selecting the best one is inherently polynomial. Moreover, if the small function class is "close" to the large function class, then the solution found in the small function class approximates the best solution in the unrestricted space of all clusterings.

We believe that the approach of using restricted function classes can be very promising, also from a practical point of view. It can be seen as a more controlled way of constructing approximate solutions of NP hard optimization problems than the standard approaches of local optimization or relaxation. While the effects of the latter cannot be controlled in general, we are able to control the effects of optimizing over smaller function classes by carefully selecting $\mathcal{F}_n$. This strategy circumvents the problem that solutions of local optimization or relaxation heuristics can be arbitrarily far away from the optimal solution.

The generic clustering algorithm we studied in this article is nearest neighbor clustering, which produces clusterings that are constant on small local neighborhoods. We have proved that this algorithm is statistically consistent for a large variety of popular clustering objective functions. Thus, as opposed to other clustering algorithms such as the K-means algorithm or spectral clustering, nearest neighbor clustering is guaranteed to converge to a minimizer of the true global optimum on the underlying space. This statement is much stronger than the results already known for K-means or spectral clustering. For K-means it has been proved that the global minimizer of the WSS objective function on the sample converges to a global minimizer on the underlying space (e.g., Pollard, 1981). However, as the standard K-means algorithm only discovers a local optimum on the discrete sample, this result does not apply to the algorithm used in practice. A related effect happens for spectral clustering, which is a relaxation attempting to minimize Ncut or RatioCut. For this class of algorithms, it has been shown that under certain conditions the solution of the relaxed problem on the finite sample converges to some limit clustering. However, this limit clustering is not necessarily the optimizer of the Ncut or RatioCut objective function.

It is interesting to note that the problems about the existing consistency results for K-means and spectral clustering are "reverse" to each other: while for K-means we know that the global minimizer converges, but this result does not apply to the algorithm used in practice, for spectral clustering there exist consistency results for the algorithm used in practice, but these results do not relate to the global minimizer. For both cases, our consistency results represent an improvement: we have constructed an algorithm which provably converges to the true limit minimizer of WSS or Ncut, respectively. The same result also holds for a large number of alternative objective functions used for clustering.

We believe that a big advantage of our approach is that both the algorithm and the statistical analysis is not restricted to center-based algorithms only, as it has been the case for most approaches in the literature (Buhmann, 1998; Ben-David, 2007; Rakhlin and Caponnetto, 2007). Instead, nearest neighbor clustering can be used as a baseline method to construct clusterings for any objective function. In von Luxburg et al. (2008) we have shown how nearest neighbor clustering can be implemented efficiently using branch and bound, and that in terms of quality, its results can compete with algorithms of spectral clustering (for the Ncut objective function) or *K*-means (for the WSS objective function). We believe that in particular for unusual objective functions for which no state of the art optimizer exists yet, nearest neighbor clustering is a promising baseline to start with. We have seen that for many commonly used objective functions, statistical guarantees for nearest neighbor clustering can be obtained, and we expect the same to be true for many more clustering objective functions.

Finally, it is a fair question how statistical consistency helps in practical applications. Is it any help in solving the big open issues in clustering, such as the question of selecting clustering algorithms for a particular data set, or selecting the number of clusters? In this generality, the answer is no. In our opinion, consistency is a *necessary* requirement which any clustering algorithm should satisfy. If an algorithm is not consistent, even with a high amount of data one cannot rely on a clustering constructed on a finite amount of data—and this is not due to computational problems, but to inherent statistical problems. Such an algorithm cannot be trusted when constructing results on a finite sample; given another sample, it might just come up with a completely different clustering. Or, the more samples one gets, the more "trivial" the solution might become (unnormalized spectral clustering is an example for such an algorithm). In this sense, consistency is just one piece of evidence to discard unreliable clustering algorithms. In our opinion, it is very hard to come up with *sufficient* conditions about "what a good clustering algorithm is". The applications of clustering are just too diverse, and 50 years of clustering literature show that people will not agree on a unique definition of what a good clustering algorithm is. This is the reason why we believe that it is very fruitful to start by studying necessary conditions first. Our current paper is meant as a contribution to this effort.

Appendix A. All Proofs

In this section we concentrate all the proofs.

A.1 The Proof of Theorem 1

The following lemma will be central in our analysis. It allows to take a supremum out of a probability.

Lemma 9 *With the notation in Theorem* 1 *we have:*

$$\mathbb{P}\big(\sup_{f\in\mathcal{F}_n}|Q_n(f)-Q(f)|\geq\varepsilon\big)\leq 2s(\widetilde{\mathcal{F}_n},2n)\frac{\sup_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q_n(f)-Q(f)|\geq\varepsilon/4)}{\inf_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q_n(f)-Q(f)|\leq\varepsilon/2)}.$$

The proof technique is similar to the one in Devroye et al. (1996), Section 12.3. The unusual term in the denominator originates in the symmetrization step. In a more standard setting where we have $\mathbb{E}Q_n = Q$, this term usually "disappears" as it can be lower bounded by $1/2$, for example using

Chebyshev's inequality (e.g., Section 12.3 of Devroye et al., 1996). Unfortunately, this does not work in our more general case, as we do not assume unbiasedness and instead also allow $\mathbb{E}Q_n \neq Q$. However, note that the ratio in Lemma 9 essentially has the form $u_n/(1-u_n)$. Thus, as soon as the term u_n in the numerator becomes non-trivial (i.e., $u_n < 1$ or say, $u_n < 3/4$), then the denominator will only play the role of a small constant (it is lower bounded by 1/4). This means that in the regime where the numerator is non-trivial, the whole bound will essentially behave like the numerator.

Proof First note that we can replace the data-dependent function class $\mathcal{F}_n$ by the class $\widetilde{\mathcal{F}_n}$ which does not depend on the data:

$$\mathbb{P}(\sup_{f\in\mathcal{F}_n}|Q_n(f)-Q(f)|\geq\varepsilon)\leq\mathbb{P}(\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q(f)|\geq\varepsilon).$$

Now we want to use a symmetrization argument. To this end, let $X'_1,\ldots,X'_n$ be a ghost sample (that is a sample drawn i.i.d. according to $\mathbb{P}$ which is independent of our first sample $X_1,\ldots,X_n$), and denote by Q'_n the empirical quality function based on the ghost sample.

Let $\widehat{f}\in\widetilde{\mathcal{F}_n}$ be such that $|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon$; if such an $\widehat{f}$ does not exist then just choose $\widehat{f}$ as some other fixed function in $\widetilde{\mathcal{F}_n}$. Note that $\widehat{f}$ is a data-dependent function depending on the sample $X_1,\ldots,X_n$. We have the following inequalities:

$$\begin{aligned}
&\mathbb{P}(\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q'_n(f)|\geq\varepsilon/2)\\
&\geq\mathbb{P}(|Q_n(\widehat{f})-Q'_n(\widehat{f})|\geq\varepsilon/2)\\
&\geq\mathbb{P}(|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon,|Q'_n(\widehat{f})-Q(\widehat{f})|\leq\varepsilon/2)\\
&=\mathbb{E}\left(\mathbb{P}(|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon,|Q'_n(\widehat{f})-Q(\widehat{f})|\leq\varepsilon/2|X_1,\ldots,X_n)\right)\\
&=\mathbb{E}\left(\mathbb{P}(|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon|X_1,\ldots,X_n)\mathbb{P}(|Q'_n(\widehat{f})-Q(\widehat{f})|\leq\varepsilon/2|X_1,\ldots,X_n)\right)\\
&=\mathbb{E}\left(1_{|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon}\mathbb{P}(|Q'_n(\widehat{f})-Q(\widehat{f})|\leq\varepsilon/2|X_1,\ldots,X_n)\right)\\
&\geq\mathbb{E}\left(1_{|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon}\inf_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q'_n(f)-Q(f)|\leq\varepsilon/2|X_1,\ldots,X_n)\right)\\
&=\mathbb{E}(1_{|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon})\mathbb{E}\left(\inf_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q'_n(f)-Q(f)|\leq\varepsilon/2|X_1,\ldots,X_n)\right)\\
&=\mathbb{P}(|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon)\inf_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q_n(f)-Q(f)|\leq\varepsilon/2)\\
&=\mathbb{P}(\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q(f)|\geq\varepsilon)\inf_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q_n(f)-Q(f)|\leq\varepsilon/2).
\end{aligned}$$

The last step is true because of the definition of $\widehat{f}$: note that due to the definition of $\widehat{f}$ the event $|Q_n(\widehat{f})-Q(\widehat{f})|\geq\varepsilon$ is true iff there exists some $f\in\widetilde{\mathcal{F}_n}$ such that $|Q_n(f)-Q(f)|\geq\varepsilon$, which is true iff $\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q(f)|\geq\varepsilon$ (recall that we assumed for ease of notations that all supremum are attained). Rearranging the inequality above leads to

$$\mathbb{P}(\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q(f)|\geq\varepsilon)\leq\frac{\mathbb{P}(\sup_{f\in\widetilde{\mathcal{F}_n}}|Q_n(f)-Q'_n(f)|\geq\varepsilon/2)}{\inf_{f\in\widetilde{\mathcal{F}_n}}\mathbb{P}(|Q_n(f)-Q(f)|\leq\varepsilon/2)}.$$

Due to the symmetrization we got rid of the quantity $Q(f)$ in the numerator. Furthermore, using the assumption of the theorem that $Q_n(f)$ does not involve any function evaluations $f(x)$ for $x \notin \{X_1, \ldots, X_n\}$ we can apply a union bound argument to move the supremum in the numerator out of the probability:

$$\begin{aligned}
&\mathbb{P}(\sup_{f \in \widetilde{\mathcal{F}_n}} |Q_n(f) - Q'_n(f)| \geq \varepsilon/2) \\
&\leq s(\widetilde{\mathcal{F}_n}, 2n) \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q'_n(f)| \geq \varepsilon/2) \\
&\leq s(\widetilde{\mathcal{F}_n}, 2n) \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| + |Q(f) - Q'_n(f)| \geq \varepsilon/2) \\
&\leq 2s(\widetilde{\mathcal{F}_n}, 2n) \sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \geq \varepsilon/4).
\end{aligned}$$

This completes the proof of the lemma. ■

Now we are ready to prove our first main theorem.

A.2 Proof of Theorem 1

Additionally to the functions f_n and f^*, we will define

$$\begin{aligned}
f_n^* &\in \operatorname*{argmin}_{f \in \mathcal{F}_n} Q(f), \\
\widetilde{f}^* &\in \operatorname*{argmin}_{f \in \mathcal{F}_n} d(f, f^*).
\end{aligned}$$

To prove the theorem we have to show that under the conditions stated, for any fixed $\varepsilon > 0$ the term $\mathbb{P}(|Q(f_n) - Q(f^*)| \geq \varepsilon)$ converges to 0. We can study each "side" of this convergence independently:

$$\mathbb{P}(|Q(f_n) - Q(f^*)| \geq \varepsilon) = \mathbb{P}(Q(f_n) - Q(f^*) \leq -\varepsilon) + \mathbb{P}(Q(f_n) - Q(f^*) \geq \varepsilon).$$

To treat the "first side" observe that if $f_n \in \mathcal{F}$ then $Q(f_n) - Q(f^*) > 0$ by the definition of f^*. This leads to

$$\mathbb{P}(Q(f_n) - Q(f^*) \leq -\varepsilon) \leq \mathbb{P}(f_n \notin \mathcal{F}).$$

Under Assumption (2) of Theorem 1 this term tends to 0.

The main work of the proof is to take care of the second side. To this end we split $Q(f_n) - Q(f^*)$ in two terms, the estimation error and the approximation error:

$$Q(f_n) - Q(f^*) = Q(f_n) - Q(f_n^*) + Q(f_n^*) - Q(f^*).$$

For a fixed $\varepsilon > 0$ we have

$$\mathbb{P}(Q(f_n) - Q(f^*) \geq \varepsilon) \leq \mathbb{P}(Q(f_n) - Q(f_n^*) \geq \varepsilon/2) + \mathbb{P}(Q(f_n^*) - Q(f^*) \geq \varepsilon/2).$$

In the following sections we will treat both parts separately.

A.2.1 ESTIMATION ERROR

The first step is to see that

$$Q(f_n) - Q(f_n^*) \leq 2 \sup_{f \in \mathcal{F}_n} |Q_n(f) - Q(f)|.$$

Indeed, since $Q_n(f_n) \leq Q_n(f_n^*)$ by the definition of f_n we have

$$\begin{aligned} Q(f_n) - Q(f_n^*) &= Q(f_n) - Q_n(f_n) + Q_n(f_n) - Q_n(f_n^*) + Q_n(f_n^*) - Q(f_n^*) \\ &\leq Q(f_n) - Q_n(f_n) + Q_n(f_n^*) - Q(f_n^*) \\ &\leq 2 \sup_{f \in \mathcal{F}_n} |Q_n(f) - Q(f)|. \end{aligned}$$

Using Lemma 9 we obtain

$$\mathbb{P}(Q(f_n) - Q(f_n^*) \geq \varepsilon/2) \leq 2s(\widetilde{\mathcal{F}_n}, 2n) \frac{\sup_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \geq \varepsilon/16)}{\inf_{f \in \widetilde{\mathcal{F}_n}} \mathbb{P}(|Q_n(f) - Q(f)| \leq \varepsilon/8)}.$$

Now observe that under Assumption (1) the numerator of the expression in the proposition tends to 0 and the denominator tends to 1, so the whole term tends to 0.

A.3 Approximation Error

By definition of f_n^* it is clear that

$$Q(f_n^*) - Q(f^*) \leq Q(\widetilde{f}^*) - Q(f^*).$$

Using Assumption (3) this leads to

$$\begin{aligned} \mathbb{P}(Q(f_n^*) - Q(f^*) \geq \varepsilon/2) &\leq \mathbb{P}(Q(\widetilde{f}^*) - Q(f^*) \geq \varepsilon/2) \\ &\leq \mathbb{P}(d(f, \widetilde{f}^*) \geq \delta(\varepsilon/2)). \end{aligned}$$

The right hand side clearly tends to 0 by Assumption (2).

■

A.4 The Proof of Theorem 2

Before proving Theorem 2, we again need to prove a few technical lemmas. The first one is a simple relation between the shattering coefficients of the nearest neighbor function classes.

Lemma 10 *Let $u \in \mathbb{N}$ and $\widetilde{\mathcal{F}_n}$ and $\widehat{\mathcal{F}_n}$ be the function sets defined in Theorem 2. Then*

$$s(\widetilde{\mathcal{F}_n}, u) \leq s(\widehat{\mathcal{F}_n}, u) \leq K^m u^{(d+1)m^2}.$$

Proof The first inequality is obvious as we have $\widetilde{\mathcal{F}_n} \subset \widehat{\mathcal{F}_n}$. For the second inequality observe that

$$s(\widehat{\mathcal{F}_n}, u) \leq K^m s^*(\widehat{\mathcal{F}_n}, u)$$

where $s^*(\widehat{\mathcal{F}_n}, u)$ is the maximal number of different ways u points can be partitioned by cells of a Voronoi partition of m points. It is well known (e.g., Section 21.5 of Devroye et al., 1996) that $s^*(\widehat{\mathcal{F}_n}, u) \leq u^{(d+1)m^2}$ for $d > 1$. Note that for $d = 1$ a similar inequality holds, we do not consider this case any further. ■

The second lemma relates a function evaluated at a point x to the same function, evaluated at the nearest neighbor of x in the training points. This lemma builds on ideas of Fritz (1975).

Lemma 11 *Let $f : \mathcal{X} \to \{1, \ldots, K\}$ be continuous almost everywhere and*

$$L_n := \mathbb{P}(f(X) \neq f(NN_m(X)) | X_1, \ldots, X_n).$$

Then for every $\varepsilon > 0$ there exists a constant $b_f(\varepsilon) > 0$ independent of n such that

$$\mathbb{P}(L_n \geq \varepsilon) \leq \frac{2}{\varepsilon} e^{-m b_f(\varepsilon)}.$$

Proof By $B(x, \delta)$ we denote the Euclidean ball of center x and radius δ. The first step of the proof consists in constructing a certain set D (depending on ε) which satisfies the following statement:

For all $\varepsilon > 0$ there exists some $\delta(\varepsilon) > 0$, a measurable set $D \subset \mathbb{R}^d$ and a constant $1 > u > 0$ such that

(a) $\mathbb{P}(D) \geq 1 - \varepsilon/2$

(b) $\forall x \in D : \mathbb{P}(B(x, \delta)) > u$

(c) $\forall x \in D$ *the function f is constant on* $B(x, \delta)$.

Assume we have such a set D. Then using Properties (c) and (a) we can see that

$$\begin{aligned} L_n &= \mathbb{P}(f(X) \neq f(NN_m(X)) | X_1, \ldots, X_n) \\ &\leq \mathbb{P}(X \notin D | X_1, \ldots, X_n) + \mathbb{P}(X \in D, |X - NN_m(X)| > \delta | X_1, \ldots, X_n) \\ &\leq \tfrac{\varepsilon}{2} + \mathbb{P}(X \in D, |X - NN_m(X)| > \delta | X_1, \ldots, X_n). \end{aligned}$$

Using the Markov inequality we can then see that

$$\begin{aligned} \mathbb{P}(L_n > \varepsilon) &\leq \mathbb{P}(\mathbb{P}(X \in D, |X - NN_m(X)| > \delta | X_1, \ldots, X_n) \geq \tfrac{\varepsilon}{2}) \\ &\leq \tfrac{2}{\varepsilon} \mathbb{E}(\mathbb{P}(X \in D, |X - NN_m(X)| > \delta | X_1, \ldots, X_n)) \\ &= \tfrac{2}{\varepsilon} \mathbb{P}(X \in D, |X - NN_m(X)| > \delta) \\ &= \tfrac{2}{\varepsilon} \int_D \mathbb{P}(|x - NN_m(x)| > \delta) \, d\mathbb{P}(x). \end{aligned}$$

Due to Property (b) we know that for all $x \in D$,

$$\begin{aligned} \mathbb{P}(|x - NN_m(x)| > \delta) &= \mathbb{P}(\forall i \in \{1, \ldots, m\}, x \notin B(X_i, \delta)) \\ &= (1 - \mathbb{P}(B(x, \delta)))^m \\ &\leq (1 - u)^m. \end{aligned}$$

Setting $b(\varepsilon) := -log(1-u) > 0$ then leads to

$$P(L_n > \varepsilon) \leq \frac{2}{\varepsilon}\mathbb{P}(D)(1-u)^m \leq \frac{2}{\varepsilon}e^{-mb(\delta(\varepsilon))}.$$

Note that this constant $b(\varepsilon)$ will also be used in several of the following lemmas. To finish the proof of the lemma we have to show how the set D can be constructed. By the assumption of the lemma we know that f is continuous a.e., and that f only takes finitely many values $1,...,K$. This implies that the set

$$C = \{x \in \mathbb{R}^d : \exists \delta > 0 : d(x,y) \leq \delta \Rightarrow f(x) = f(y)\}$$

satisfies $\mathbb{P}(C) = 1$. Furthermore, for any $\delta > 0$ we define the set

$$A_\delta = \{x \in C : d(x,y) \leq \delta \Rightarrow f(x) = f(y)\}.$$

We have $\cup_\delta A_\delta = C$, and for $\sigma > \delta$ we have $A_\sigma \subset A_\delta$. This implies that given some $\varepsilon > 0$ there exists some $\delta(\varepsilon) > 0$ such that $\mathbb{P}(A_{\delta(\varepsilon)}) \geq 1 - \varepsilon/4$. By construction, all points in $A_{\delta(\varepsilon)}$ satisfy Property (c).

As the next step, we can see that for every $\delta > 0$ one has $\mathbb{P}(B(x,\delta)) > 0$ almost surely (with respect to x). Indeed, the set $U = \{x : \exists \delta > 0 : \mathbb{P}(B(x,\delta)) = 0\}$ is a union of sets of probability zero. So using the fact that $\mathbb{R}^d$ is separable we see that $\mathbb{P}(U) = 0$. Thus, $\mathbb{P}(\mathbb{P}(B(X,\delta)|X) > 0) = 1$, which implies $\mathbb{P}(\mathbb{P}(B(X,\delta)|X) > \frac{1}{n}) \to 1$. This means that given $\varepsilon > 0$ and $\delta > 0$ there exists a set A and a constant $u > 0$ such that $\mathbb{P}(A) \geq 1 - \varepsilon/4$ and $\forall x \in A, \mathbb{P}(B(x,\delta)) > u$. So all points in A satisfy Property (b).

Now finally define the set $D = A \bigcap A_{\delta(\varepsilon)}$. By construction, this set has probability $P(D) \geq \varepsilon/2$, so it satisfies Property (a). It satisfies Properties (b) and (c) by construction of A and $A_{\delta(\varepsilon)}$, respectively. ■

A.5 Proof of Theorem 2

To prove this theorem we will verify that the conditions (1) - (3) of Theorem 1 are satisfied for the function classes studied in Theorem 2.

Lemma 10 proves that Condition (1) of Theorem 2 implies Condition (1) of Theorem 1. Moreover, it is obvious that Condition (3) of Theorem 2 implies Condition (3) of Theorem 1.

Thus we only have to prove Condition (2) of Theorem 1. We begin by proving that $\mathbb{P}(f_n \notin \mathcal{F}) \to 0$. As $f_n \in \mathcal{F}_n$ by definition we have that $\Phi_n(f_{n,k}) > a_n$ for all $k = 1,\ldots,K$. A union bound argument shows that

$$\mathbb{P}(f_n \notin \mathcal{F}) \leq K \sup_k \mathbb{P}(\Phi(f_{n,k}) \leq a).$$

Using the same techniques as in the proof of Lemma 9 we can see that

$$\begin{aligned}
\mathbb{P}(\Phi(f_{n,k}) \leq a) &\leq \mathbb{P}(\Phi_n(f_{n,k}) - \Phi(f_{n,k}) \geq a_n - a) \\
&\leq \mathbb{P}(\sup_{g \in \mathcal{F}_n} \Phi_n(g_k) - \Phi(g_k) \geq a_n - a) \\
&\leq 2s(\widehat{\mathcal{F}_n}, 2n) \frac{\sup_{g \in \widehat{\mathcal{F}_n}} \mathbb{P}(\Phi_n(g_k) - \Phi(g_k) \geq (a_n - a)/4)}{\inf_{g \in \widehat{\mathcal{F}_n}} \mathbb{P}(\Phi_n(g_k) - \Phi(g_k) \leq (a_n - a)/2)}.
\end{aligned}$$

Moreover, we already proved in Lemma 10 that $s(\widehat{\mathcal{F}_n}, 2n) \leq K^m (2n)^{(d+1)m^2}$. Condition (5) of Theorem 2 then implies that $\mathbb{P}(\Phi(f_{n,k}) \leq a)$ tends to 0.

Now we have to prove that for $f \in \mathcal{F}$ the term $d(f, \mathcal{F}_n) := \min_{g \in \mathcal{F}_n} d(f,g)$ tends to 0 in probability. Let $\widetilde{f}(x) = f(NN_m(x))$. If $\widetilde{f} \in \mathcal{F}_n$ then $d(f, \mathcal{F}_n) \leq d(f, \widetilde{f})$, so the following holds true:

$$\mathbb{P}(d(f, \mathcal{F}_n) \geq \varepsilon) \leq \mathbb{P}(\widetilde{f} \notin \mathcal{F}_n) + \mathbb{P}(d(f, \widetilde{f}) \geq \varepsilon).$$

The second term on the right hand side tends to 0 because of Lemma 11. To deal with the first term on the right hand side, observe that

$$\mathbb{P}(\widetilde{f} \notin \mathcal{F}_n) \leq K \sup_k \mathbb{P}(\Phi_n(\widetilde{f}_k) \leq a_n).$$

Because of Condition (4), for all $\varepsilon > 0$, $f \in \mathcal{F}$ and $g \in \widehat{\mathcal{F}_n}$ there exists $\delta(\varepsilon) > 0$ such that

$$d(f,g) \leq \delta(\varepsilon) \Rightarrow \Phi(f_k) - \Phi(g_k) \leq \varepsilon.$$

Define $a_n^f := \inf_k \Phi(f_k) - a_n$. Since $f \in \mathcal{F}$ there exists N such that $n \geq N \Rightarrow a_n^f > 0$. For $n \geq N$ we have the following inequalities:

$$\begin{aligned}
&\mathbb{P}(\Phi_n(\widetilde{f}_k) \leq a_n) \\
&= \mathbb{P}(\Phi(f_k) - \Phi_n(\widetilde{f}_k) \geq \Phi(f_k) - a_n) \\
&= \mathbb{P}(\Phi(f_k) - \Phi(\widetilde{f}_k) + \Phi(\widetilde{f}_k) - \Phi_n(\widetilde{f}_k) \geq \Phi(f_k) - a_n) \\
&\leq \mathbb{P}(\Phi(f_k) - \Phi(\widetilde{f}_k) \geq (\Phi(f_k) - a_n)/2) + \mathbb{P}(\Phi(\widetilde{f}_k) - \Phi_n(\widetilde{f}_k) \geq (\Phi(f_k) - a_n)/2) \\
&\leq \mathbb{P}(\Phi(f_k) - \Phi(\widetilde{f}_k) \geq a_n^f/2) + \mathbb{P}(\Phi(\widetilde{f}_k) - \Phi_n(\widetilde{f}_k) \geq a_n^f/2) \\
&\leq \mathbb{P}(d(f, \widetilde{f}) > \delta(a_n^f/2)) + \mathbb{P}(\sup_{g \in \widehat{\mathcal{F}_n}} \Phi(g_k) - \Phi_n(g_k) \geq a_n^f/2) \\
&\leq \tfrac{2}{\delta(a_n^f/2)} e^{-mb(\delta(a_n^f/2))} + \mathbb{P}(\sup_{g \in \widehat{\mathcal{F}_n}} \Phi(g_k) - \Phi_n(g_k) \geq a_n^f/2).
\end{aligned}$$

If $m \to \infty$ then the first term goes to 0. Indeed, $\delta(a_n^f/2)$ and $b(\delta(a_n^f/2))$ tend to positive constants since $f \in \mathcal{F}$ and thus $a_n^f \to \inf_k \Phi(f_k) - a > 0$. For the second term, the key step is to see that by the techniques used in the proof of Lemma 9 we get

$$\begin{aligned}
&\mathbb{P}(\sup_{g \in \widehat{\mathcal{F}_n}} \Phi(g_k) - \Phi_n(g_k) \geq a_n^f/2) \\
&\leq 2K^m (2n)^{(d+1)m^2} \frac{\sup_{g \in \widehat{\mathcal{F}_n}} \mathbb{P}(\Phi(g_k) - \Phi_n(g_k) \geq a_n^f/8)}{\inf_{g \in \widehat{\mathcal{F}_n}} \mathbb{P}(\Phi(g_k) - \Phi_n(g_k) \leq a_n^f/4)}.
\end{aligned}$$

Under Condition (2) this term tends to 0. ■

A.6 The Proofs of the Consistency Theorems 3, 5, 7 and 8

All these theorems are applications of Theorem 2 to specific objective functions Q_n and Q and to specific functions Φ_n and Φ. For all of them, we individually have to check whether the conditions in Theorem 2 are satisfied. In this section, we do not follow the order of the Theorems in the paper. This is only due to better readability of the proofs.

In most of these proves, we will use the McDiarmid inequality (McDiarmid, 1989), which we recall for the convenience of the reader:

Theorem 12 (McDiarmid inequality) *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent random variables. Let $g : (\mathbb{R}^d)^n \to \mathbb{R}$ be measurable and $c > 0$ a constant such that for all $1 \leq i \leq n$ we have*

$$\sup_{x_1,\ldots,x_n,x'\in\mathbb{R}^d} g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x',x_{i+1},\ldots,x_n) \leq c.$$

Then

$$\mathbb{P}(|g(X_1,\ldots,X_n) - \mathbb{E}g(X_1,\ldots,X_n)| \geq \varepsilon) \leq 2e^{-\frac{2\varepsilon^2}{nc^2}}.$$

Moreover, several times we will use the fact that $a_n \to a, m \to \infty$ and $\frac{m^2 \log n}{n(a-a_n)^2} \to 0$ implies that $n(a-a_n)^2 \to \infty$ and $\frac{m^2 \log n}{n} \to 0$.

Before we look at the "combined" objective functions such as Ncut, RatioCut, WSS, we will prove some technical conditions about their "ingredients" cut, vol, $E\, f_k(X)$ and WS.

Lemma 13 (Conditions (2), (4), and (5) for cut**,** vol**,** $\mathbb{E}\, f_k(X)$**, and** WS**)** *Assume that*

$$\frac{m^2 \log n}{n(a-a_n)^2} \to 0$$

then vol, cut, $\mathbb{E}\, f_k(X)$ *and* WS *satisfy Conditions (2), (4) and (5) of Theorem 2.*

Proof To prove Conditions (2) and (5) we are going to use the McDiarmid inequality. Observe that if one replaces one variable X_i by a new one X_i', then vol_n changes by at most $2C/n$, cut_n changes by at most $2C/n$, $\mathrm{WS}(f_k)$ changes by at most $2C/n$, and $n_k(f)$ changes by at most $1/n$. Using the McDiarmid inequality, this implies that for all $g \in \widehat{\mathcal{F}_n}$ and $\varepsilon > 0$

$$\begin{aligned}
&\mathbb{P}(|\mathrm{vol}_n(g_k) - \mathrm{vol}(g_k)| \geq \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{2C^2}},\\
&\mathbb{P}(|\mathrm{cut}_n(g_k) - \mathrm{cut}(g_k)| \geq \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{2C^2}},\\
&\mathbb{P}(|\mathrm{WS}_n(g_k) - \mathrm{WS}(g_k)| \geq \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{2C^2}},\\
&\mathbb{P}(|n_k(g) - \mathbb{E}g_k(X)| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}.
\end{aligned}$$

So to prove Condition (2) we have to show that

$$\forall \varepsilon > 0, K^m (2n)^{(d+1)m^2} e^{-n\varepsilon} \to 0.$$

This follows clearly from $K^m (2n)^{(d+1)m^2} e^{-n\varepsilon} = e^{-n\left(\frac{m\log K + (d+1)m^2 \log(2n)}{-n} + \varepsilon\right)}$ and $\frac{m^2 \log n}{n} \to 0$. Moreover, since $n(a-a_n)^2 \to \infty$ Condition (5) is also true.

To prove (4) for each of the objective functions, let $f, g \in \mathcal{H}$ and f_k and g_k be the corresponding cluster indicator functions for cluster k. Then we can see that

$$\begin{aligned}
|\operatorname{vol}(g_k) - \operatorname{vol}(f_k)| &= |\int\int (f_k(X) - g_k(X)) s(X,Y)\, dP(X) dP(Y)| \\
&\leq C \int_{\{f_k = g_k\}} 0\, dP(X) dP(Y) + C \int_{\{f_k = g_k\}^c} 1\, dP(X) dP(Y) \\
&= C\mathbb{P}(f_k \neq g_k) \\
&\leq C d(f,g),
\end{aligned}$$

$$\begin{aligned}
|\operatorname{cut}(g_k) - \operatorname{cut}(f_k)| &= |\int\int f_k(X)(1 - f_k(Y)) s(X,Y) - g_k(X)(1 - g_k(Y)) s(X,Y)\, dP(X) dP(Y)\,| \\
&\leq C \int\int_{\{f=g\}^2} 0\, dP(X) dP(Y) + C \int\int_{(\{f=g\}^2)^c} 1\, dP(X) dP(Y) \\
&= C(1 - \mathbb{P}(f(X) = g(X))^2) \\
&= C(1 - (1 - d(f,g))^2) \\
&\leq 2C d(f,g),
\end{aligned}$$

$$|\mathbb{E} f_k(X) - \mathbb{E} g_k(X)| \leq d(f,g),$$

$$\begin{aligned}
|\operatorname{WS}(f_k) - \operatorname{WS}(g_k)| &= |\int\int (f_k(X) f_k(Y) - g_k(X) g_k(Y)) s(X,Y)\, dP(X) dP(Y)\,| \\
&\leq \int_{\{f=g\}^2} 0\, dP(X) dP(Y) + C \int_{(\{f=g\}^2)^c} 1\, dP(X) dP(Y) \\
&= C(1 - \mathbb{P}(f = g)^2) \\
&= C(1 - (1 - d(f,g))^2) \\
&\leq 2C d(f,g).
\end{aligned}$$

■

Now we are going to check that the "combined" objective functions Ncut, RatioCut, Mod, WSS, BWR satisfy the conditions of Theorem 2. For many of the objective functions, one important step in the proof is to separate the convergence of the whole term into the convergence of the numerator and the denominator.

Lemma 14 (Condition (1) for Ncut**)** *Assume that*

$$\frac{m^2 \log n}{n} \to 0$$

then Ncut *satisfies Condition (1) of Theorem* 2.

Proof We first want to split the deviations of Ncut into the ones of cut and vol, respectively. To this end we want to show that for any $f \in \widetilde{\mathcal{F}_n}$

$$\{|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| \leq \tfrac{a}{2}\varepsilon\} \bigcap \{|\operatorname{vol}_n(f_k) - \operatorname{vol}(f_k)| \leq \tfrac{a}{2}\varepsilon\}$$

$$\subset \{|\tfrac{\operatorname{cut}_n(f_k)}{\operatorname{vol}_n(f_k)} - \tfrac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)}| \leq \varepsilon\}.$$

This can be seen as follows. Assume that $|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| \leq \varepsilon$ and $|\operatorname{vol}_n(f_k) - \operatorname{vol}(f_k)| \leq \varepsilon$. If $\operatorname{vol}(f_k) \neq 0$ then we have (using the facts that $\operatorname{cut}(f_k) \leq \operatorname{vol}(f_k)$ and that $\operatorname{vol}_n(f_k) > a_n > a$ by definition of $\widetilde{\mathcal{F}_n}$):

$$\begin{aligned}
\tfrac{\operatorname{cut}_n(f_k)}{\operatorname{vol}_n(f_k)} - \tfrac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)} &= \tfrac{\operatorname{cut}_n(f_k)\operatorname{vol}(f_k) - \operatorname{cut}(f_k)\operatorname{vol}_n(f_k)}{\operatorname{vol}_n(f_k)\operatorname{vol}(f_k)} \\
&\leq \tfrac{(\operatorname{cut}(f_k)+\varepsilon)\operatorname{vol}(f_k) - \operatorname{cut}(f_k)(\operatorname{vol}(f_k)-\varepsilon)}{\operatorname{vol}_n(f_k)\operatorname{vol}(f_k)} \\
&= \tfrac{\varepsilon}{\operatorname{vol}_n(f_k)} \tfrac{\operatorname{cut}(f_k)+\operatorname{vol}(f_k)}{\operatorname{vol}(f_k)} \\
&\leq \tfrac{2\varepsilon}{a}.
\end{aligned}$$

On the other hand, if $\operatorname{vol}(f_k) = 0$ then we have $\operatorname{cut}(f_k) = 0$, which implies $\operatorname{cut}_n(f_k) \leq \varepsilon$ by the assumption above. Thus the following statement holds true:

$$\frac{\operatorname{cut}_n(f)}{\operatorname{vol}_n(f)} - \frac{\operatorname{cut}(f)}{\operatorname{vol}(f)} = \frac{\operatorname{cut}_n(f)}{\operatorname{vol}_n(f)} \leq \frac{\varepsilon}{a} \leq \frac{2\varepsilon}{a}.$$

Using the same technique we have the same bound for $\frac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)} - \frac{\operatorname{cut}_n(f_k)}{\operatorname{vol}_n(f_k)}$, which proves our set inclusion.

Now we apply a union bound and the McDiarmid inequality. For the latter, note that if one changes one X_i then $\operatorname{cut}_n(f)$ and $\operatorname{vol}_n(f)$ will change at most by $2C/n$. Together all this leads to

$$\begin{aligned}
&\mathbb{P}(|\operatorname{Ncut}(f) - \operatorname{Ncut}_n(f)| > \varepsilon) \\
&\leq K \sup_k \mathbb{P}(|\frac{\operatorname{cut}_n(f_k)}{\operatorname{vol}_n(f_k)} - \frac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)}| > \varepsilon/K) \\
&\leq K \sup_k \Big(\mathbb{P}(|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| > \frac{a}{2K}\varepsilon) + \mathbb{P}(|\operatorname{vol}_n(f_k) - \operatorname{vol}(f_k)| > \frac{a}{2K}\varepsilon) \Big) \\
&\leq 4Ke^{-\frac{na^2\varepsilon^2}{8C^2K^2}}.
\end{aligned}$$

To finish we have to prove that

$$\forall \varepsilon > 0, K^{m+1}(2n)^{(d+1)m^2} e^{-n\varepsilon} \to 0.$$

This follows clearly from $K^{m+1}(2n)^{(d+1)m^2} e^{-n\varepsilon} = e^{-n\left(\frac{(m+1)\log K + (d+1)m^2\log(2n)}{-n} + \varepsilon\right)}$ and $\frac{m^2 \log n}{n} \to 0$. ■

Lemma 15 (Condition (3) for Ncut**)** Ncut *satisfies Condition (3) of Theorem* 2.

Proof Let $f \in \mathcal{F}, g \in \widetilde{\mathcal{F}_n}$. In the proof of Lemma 13 we have already seen that

$$\begin{aligned} |\operatorname{cut}(f_k) - \operatorname{cut}(g_k)| &\leq 2Cd(f,g), \\ |\operatorname{vol}(f_k) - \operatorname{vol}(g_k)| &\leq 2Cd(f,g). \end{aligned}$$

If $\operatorname{vol}(g) \neq 0$ then we have (using the fact that we always have $\operatorname{cut}(f) \leq \operatorname{vol}(f)$):

$$\begin{aligned} \tfrac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)} - \tfrac{\operatorname{cut}(g_k)}{\operatorname{vol}(g_k)} &= \tfrac{\operatorname{cut}(f_k)\operatorname{vol}(g_k) - \operatorname{cut}(g_k)\operatorname{vol}(f_k)}{\operatorname{vol}(f_k)\operatorname{vol}(g_k)} \\ &\leq \tfrac{(\operatorname{cut}(g_k)+2Cd(f,g))\operatorname{vol}(\widetilde{f}) - \operatorname{cut}(g_k)(\operatorname{vol}(g_k)-2Cd(f,g))}{\operatorname{vol}(f_k)\operatorname{vol}(g_k)} \\ &= \tfrac{2Cd(f,g)}{\operatorname{vol}(f_k)} \tfrac{\operatorname{vol}(g_k)+\operatorname{cut}(g_k)}{\operatorname{vol}(g_k)} \\ &\leq \tfrac{4C}{a} d(f,g)). \end{aligned}$$

On the other hand if $\operatorname{vol}(g_k) = 0$ then we have $|\operatorname{cut}(f_k)| \leq |\operatorname{vol}(f_k)| \leq 2Cd(f,g)$, in which case the following holds true:

$$\frac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)} - \frac{\operatorname{cut}(g_k)}{\operatorname{vol}(\widetilde{f_k})} = \frac{\operatorname{cut}(f_k)}{\operatorname{vol}(f_k)} \leq \frac{2Cd(f,g)}{a} \leq \frac{4C}{a} d(f,g).$$

So all in all we have

$$\operatorname{Ncut}(f) - \operatorname{Ncut}(g) \leq \frac{4CK}{a} d(f,g).$$

We can use the same technique to bound $\operatorname{Ncut}(g) - \operatorname{Ncut}(f)$. This proves that Ncut is Lipschitz and thus uniformly continuous. ■

Lemma 16 (Condition (1) for RatioCut**)** *Assume that*

$$\frac{m^2 \log n}{n} \to 0$$

then RatioCut *satisfies Condition* (1) *of Theorem* 2.

Proof Using exactly the same proof as for Lemma 14 (just changing $\operatorname{vol}_n(f_k)$ to n_k and $\operatorname{vol}(f_k)$ to $\mathbb{E} f_k(X)$ and using the fact that $\operatorname{cut}(f_k) \leq C\mathbb{E} f_k(X)$) we get

$$\begin{aligned} &\mathbb{P}(|\operatorname{RatioCut}_n(f) - \operatorname{RatioCut}(f)| > \varepsilon) \\ &\leq K \sup_k \left(\mathbb{P}(|\operatorname{cut}_n(f_k) - \operatorname{cut}(f_k)| > \frac{a}{(S+1)K}\varepsilon) + \mathbb{P}(|n_k(f) - \mathbb{E} f_k(X)| > \frac{a}{(S+1)K}\varepsilon) \right). \end{aligned}$$

Now a simple McDiarmid argument (using again the fact that changing one X_i changes cut_n by at most $2S/n$) gives

$$\mathbb{P}(|\operatorname{RatioCut}_n(f) - \operatorname{RatioCut}(f)| > \varepsilon) \leq 2Ke^{-\frac{na^2\varepsilon}{8C^2K^2}} + 2Ke^{-\frac{na^2\varepsilon^2}{2K^2}}.$$

We conclude the proof with the same argument as in Lemma 14. ■

Lemma 17 (Condition (3) for RatioCut**)** RatioCut *satisfies Condition (3) of Theorem* 2.

Proof This follows by the same proof as Lemma 14, just changing $\mathrm{vol}_n(f_k)$ to n_k, $\mathrm{vol}(f_k)$ to $\mathbb{E} f_k(X)$ and using the fact that $\mathrm{cut}(f_k) \leq C\mathbb{E} f_k(X)$. ■

Lemma 18 (Condition (1) for BWR**)** *If* $m^2 \log n/n \to 0$*, then* BWR *satisfies Condition (1) of Theorem* 2.

Proof Let $f \in \widetilde{\mathcal{F}_n}$. Let $\varepsilon \leq a/2$. If $|\mathrm{WS}_n(f_k) - \mathrm{WS}(f_k)| \leq \varepsilon$ and $|\mathrm{cut}_n(f_k) - \mathrm{cut}(f_k)| \leq \varepsilon$ then $\mathrm{WS}(f_k) \geq a/2 > 0$ (because $\mathrm{WS}_n(f_k) > a_n > a$ since $f \in \widetilde{\mathcal{F}_n}$). This implies

$$\begin{aligned}
\frac{\mathrm{cut}_n(f_k)}{\mathrm{WS}_n(f_k)} - \frac{\mathrm{cut}(f_k)}{\mathrm{WS}(f_k)} &= \frac{\mathrm{WS}(f_k)\,\mathrm{cut}_n(f_k) - \mathrm{WS}_n(f_k)\,\mathrm{cut}(f_k)}{\mathrm{WS}_n(f_k)\,\mathrm{WS}(f_k)} \\
&\leq \frac{\mathrm{WS}(f_k)(\mathrm{cut}(f_k)+\varepsilon) - (\mathrm{WS}(f_k)-\varepsilon)\,\mathrm{cut}(f_k)}{\mathrm{WS}_n(f_k)\,\mathrm{WS}(f_k)} \\
&= \frac{\varepsilon}{\mathrm{WS}_n(f_k)} \frac{\mathrm{WS}(f_k)+\mathrm{cut}(f_k)}{\mathrm{WS}(f_k)} \\
&\leq \frac{2C\varepsilon}{a^2}.
\end{aligned}$$

The analogous statement holds for $\frac{\mathrm{cut}(f_k)}{\mathrm{WS}(f_k)} - \frac{\mathrm{cut}_n(f_k)}{\mathrm{WS}_n(f_k)}$. Thus, if $\varepsilon \leq C/a$ then

$$\begin{aligned}
&\{|\mathrm{WS}_n(f_k) - \mathrm{WS}(f_k)| \leq a^2\varepsilon/(2C)\} \cap \{|\mathrm{cut}_n(f_k) - \mathrm{cut}(f_k)| \leq a^2\varepsilon/(2C)\} \\
&\subset \left\{\left|\frac{\mathrm{cut}_n(f_k)}{\mathrm{WS}_n(f_k)} - \frac{\mathrm{cut}(f_k)}{\mathrm{WS}(f_k)}\right| \leq \varepsilon\right\}.
\end{aligned}$$

As a consequence, if $\varepsilon \leq CK/a$ we have

$$\begin{aligned}
\mathbb{P}(|\mathrm{BWR}_n(f) - \mathrm{BWR}(f)| > \varepsilon) &\leq K \sup_k \mathbb{P}\left(\left|\frac{\mathrm{cut}_n(f_k)}{\mathrm{WS}_n(f_k)} - \frac{\mathrm{cut}(f_k)}{\mathrm{WS}(f_k)}\right| > \varepsilon/K\right) \\
&\leq K \sup_k \left(\mathbb{P}(|\mathrm{WS}_n(f_k) - \mathrm{WS}(f_k)| > a^2\varepsilon/(2CK)) + \mathbb{P}(|\mathrm{cut}_n(f_k) - \mathrm{cut}(f_k)| > a^2\varepsilon/(2CK))\right).
\end{aligned}$$

Using the McDiarmid inequality together with the fact that changing one point changes cut_n and WS_n by at most $C/(2n)$, we get for $\varepsilon \leq CK/a$:

$$\mathbb{P}(|\mathrm{BWR}_n(f) - \mathrm{BWR}(f)| > \varepsilon) \leq 4Ke^{-\frac{na^4\varepsilon^2}{8C^4K^2}}.$$

On the other hand, for $\varepsilon > CK/a$ we have

$$\begin{aligned}
&\mathbb{P}(|\mathrm{BWR}_n(f) - \mathrm{BWR}(f)| > \varepsilon) \\
&\leq \mathbb{P}(|\mathrm{BWR}_n(f) - \mathrm{BWR}(f)| > SK/a) \\
&\leq 4Ke^{-\frac{na^4(SK/a)^2}{8C^4K^2}}.
\end{aligned}$$

So all in all we have proved that

$$\mathbb{P}(|\mathrm{BWR}_n(f)-\mathrm{BWR}(f)|>\varepsilon)\leq 2Ke^{-\frac{na^4(\min(\varepsilon,CK/a))^2}{8C^4K^2}}.$$

We conclude the proof with the same argument as in Lemma 14. ■

Lemma 19 (Condition (3) for BWR) BWR *satisfies Condition (3) of Theorem* 2.

Proof Let $\varepsilon>0$, $f\in\mathcal{F}$ and $g\in\widetilde{\mathcal{F}_n}$. We have already proved the two following inequalities (in the proofs of Lemmas 13 and 15):

$$|\mathrm{cut}(f_k)-\mathrm{cut}(g_k)|\leq 2Cd(f,g),$$

$$|\mathrm{WS}(f_k)-\mathrm{WS}(g_k)|\leq 2Cd(f,g).$$

If $2Cd(f,g)\leq a/2$, then using that $\mathrm{WS}(f_k)>a$ we get $\mathrm{WS}(g_k)\geq a/2>0$. By the same technique as at the beginning of Lemma 18 we get

$$|\mathrm{BWR}(f)-\mathrm{BWR}(g)|\leq\frac{2CK}{a^2}2Cd(f,g).$$

Written a bit differently,

$$d(f,g)\leq\frac{a}{4C}\Rightarrow|\mathrm{BWR}(f)-\mathrm{BWR}(g)|\leq\frac{4C^2K}{a^2}d(f,g).$$

Now recall that we want to prove that there exists $\delta>0$ such that $d(f,g)\leq\delta\Rightarrow|\mathrm{BWR}(f)-\mathrm{BWR}(g)|\leq\varepsilon$.

If $\varepsilon\leq CK/a$ then we have:

$$d(f,g)\leq\frac{a^2}{4C^2K}\varepsilon\leq\frac{a}{4C}\Rightarrow|\mathrm{BWR}(f)-\mathrm{BWR}(g)|\leq\frac{4C^2K}{a^2}d(f,g)\leq\varepsilon.$$

On the other hand, if $\varepsilon>CK/a$ then

$$d(f,g)\leq\frac{a}{4C}\Rightarrow|\mathrm{BWR}(f)-\mathrm{BWR}(g)|\leq\frac{4C^2K}{a^2}d(f,g)\leq CK/a\leq\varepsilon$$

so we have proved the lemma. ■

Lemma 20 (Condition (1) for WSS) *If* $\frac{m^2\log n}{n}\to 0$ *and that* $\mathrm{supp}\,\mathbb{P}\subset B(0,A)$, *then* WSS *satisfies Condition (1) of Theorem* 2.

Proof Let $f\in\widetilde{\mathcal{F}_n}$. First note that

$$\begin{aligned}|\mathrm{WSS}_n(f)-\mathrm{WSS}(f)|&=\left|\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\|X_i-c_{k,n}\|^2-\mathbb{E}\sum_{k=1}^K f_k(X)\|X-c_k\|^2\right|\\&\leq\left|\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\|X_i-c_{k,n}\|^2-\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\|X_i-c_k\|^2\right|\\&\quad+\left|\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\|X_i-c_k\|^2-\mathbb{E}\sum_{k=1}^K f_k(X)\|X-c_k\|^2\right|.\end{aligned}$$

Now we will bound the probability for each of the terms on the right hand side. For the second term we can simply apply McDiarmid's inequality. Due to the assumption that $\operatorname{supp}\mathbb{P} \subset B(0,A)$ we know that for any two points $x,y \in \operatorname{supp}\mathbb{P}$ we have $\|x-y\| \leq 2A$. Thus if one changes one variable X_i then the term $\frac{1}{n}\sum_{i=1}^n \sum_{k=1}^K f_k(X_i)\|X_i - c_k\|^2$ will change by at most $A^2/(4n)$. This leads to

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\|X_i-c_k\|^2 - \mathbb{E}\sum_{k=1}^K f_k(X)\|X-c_k\|^2\right| \geq \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{A^4}}.$$

Now we have to take care of the first term, which can be written as

$$\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\left(\|X_i-c_{k,n}\|^2 - \|X_i-c_k\|^2\right).$$

The triangle inequality gives

$$\|X_i-c_{k,n}\|^2 \leq (\|X_i-c_k\| + \|c_{k,n}-c_k\|)^2,$$

and together with the fact that $\operatorname{supp}\mathbb{P} \subset B(0,A)$ this leads to

$$\|X_i-c_{k,n}\|^2 - \|X_i-c_k\|^2 \leq 6A\|c_{k,n}-c_k\|.$$

So at this point we have

$$\left|\frac{1}{n}\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\|X_i-c_{k,n}\|^2 - \|X_i-c_k\|^2\right| \leq 6A\sup_k\|c_{k,n}-c_k\|.$$

We will denote the j-th coordinate of a vector X by X^j. Recall that d denotes the dimensionality of our space. Using this notation we have

$$\|c_{k,n}-c_k\|^2 = \sum_{j=1}^d\left(\frac{\mathbb{E}\,f_k(X)X^j}{\mathbb{E}\,f_k(X)} - \frac{1}{n_k}\frac{1}{n}\sum_{i=1}^n f_k(X_i)X_i^j\right)^2.$$

Our goal will be to apply the McDiarmid inequality to each coordinate. Before we can do this, we want to show that

$$\{|n_k - \mathbb{E}\,f_k(X)| \leq \tfrac{a\varepsilon}{A+1}\} \cap \{|\tfrac{1}{n}\sum_{i=1}^n f_k(X_i)X_i^j - \mathbb{E}\,f_k(X)X^j| \leq \frac{a\varepsilon}{A+1}\} \subset \{|c_k^j - c_{k,n}^j| \leq \varepsilon\}.$$

To this end, assume that $|n_k - \mathbb{E}\,f_k(X)| \leq \varepsilon$ and $|\frac{1}{n}\sum_{i=1}^n f_k(X_i)X_i^j - \mathbb{E}\,f_k(X)X^j| \leq \varepsilon$.

In case $\mathbb{E}\,f_k(X) \neq 0$ we have

$$\begin{aligned}
c_k^j - c_{k,n}^j &= \frac{n_k\mathbb{E}\,f_k(X)X^j - \mathbb{E}\,f_k(X)\frac{1}{n_k}\frac{1}{n}\sum_{i=1}^n f_k(X_i)X_i^j}{n_k\mathbb{E}\,f_k(X)} \\
&\leq \frac{(\mathbb{E}\,f_k(X)+\varepsilon)\mathbb{E}\,f_k(X)X^j - \mathbb{E}\,f_k(X)(\mathbb{E}\,f_k(X)X^j-\varepsilon)}{n_k\mathbb{E}\,f_k(X)} \\
&= \frac{\varepsilon}{n_k}\frac{\mathbb{E}\,f_k(X)X^j + \mathbb{E}\,f_k(X)}{\mathbb{E}\,f_k(X)} \\
&\leq \frac{(A+1)\varepsilon}{a}
\end{aligned}$$

and similarly for $c^j_{k,n} - c^j_k$.

On the other hand, in case $\mathbb{E} f_k(X) = 0$ we also have $\mathbb{E} f_k(X)X^j = 0$ (as f_k is a non-negative function and $|X|$ is bounded by A). Together with the assumption this means that $\frac{1}{n}\sum_{i=1}^n f_k(X_i)X_i^j \leq \varepsilon$. This implies

$$|c^j_k - c^j_{k,n}| = \frac{1}{n_k}\frac{1}{n}\sum_{i=1}^n f_k(X_i)X_i^j \leq \frac{\varepsilon}{a} \leq \frac{(A+1)\varepsilon}{a}$$

which shows the inclusion stated above. The McDiarmid inequality now yields the two statements

$$\begin{aligned}
&\mathbb{P}(|n_k - \mathbb{E} f_k(X)| > \varepsilon) \leq 2e^{-2n\varepsilon^2},\\
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n f_k(X_i)X_i^j - \mathbb{E} f_k(X)X^j\right| > \varepsilon\right) \leq 2e^{-\frac{2n\varepsilon^2}{A^2}}.
\end{aligned}$$

Together they show that for the coordinate-wise differences

$$\mathbb{P}(|c^j_k - c^j_{k,n}| > \varepsilon) \leq 2e^{-\frac{2na^2\varepsilon^2}{(A+1)^2}} + 2e^{-\frac{2na^2\varepsilon^2}{A^2(A+1)^2}} \leq 4e^{-\frac{2na^2\varepsilon^2}{\max(1,A^2)(A+1)^2}}.$$

This leads to

$$\begin{aligned}
\mathbb{P}(\|c_k - c_{k,n}\| > \varepsilon) &= \mathbb{P}(\sum_{j=1}^d |c^j_k - c^j_{k,n}|^2 > \varepsilon^2) \leq d \sup_j \mathbb{P}(|c^j_k - c^j_{k,n}| > \varepsilon/\sqrt{d})\\
&\leq 4de^{-\frac{2na^2\varepsilon^2}{d\max(1,A^2)(A+1)^2}}.
\end{aligned}$$

Combining all this leads to a bound for the first term of the beginning of the proof:

$$\begin{aligned}
&\mathbb{P}\left(\left|\tfrac{1}{n}\textstyle\sum_{i=1}^n\sum_{k=1}^K f_k(X_i)\left(\|X_i - c_{k,n}\|^2 - \|X_i - c_k\|^2\right)\right| \geq \varepsilon\right)\\
&\leq \mathbb{P}\left(\sup_k \|c_{k,n} - c_k\| \geq \varepsilon/(6A)\right)\\
&\leq K \sup_k \mathbb{P}(\|c_{k,n} - c_k\| \geq \varepsilon/(6A))\\
&\leq 4dKe^{-\frac{na^2\varepsilon^2}{18d\max(1,A^2)A^2(A+1)^2}}.
\end{aligned}$$

Now we combine the probabilities for the first and the second term from the beginning of the proof using a union bound to get

$$\mathbb{P}(|\mathrm{WSS}_n(f) - \mathrm{WSS}(f)| > \varepsilon) \leq 4dKe^{-\frac{na^2\varepsilon}{18d\max(1,A^2)A^2(A+1)^2}} + 2e^{-\frac{8n\varepsilon^2}{A^4}}.$$

We conclude the proof with the same argument as in Lemma 14. ■

Lemma 21 (Condition (3) for WSS) *Assume that* $\operatorname{supp}\mathbb{P} \subset B(0,A)$ *then* WSS *satisfies Condition (3) of Theorem* 2.

Proof Let $f \in \mathcal{F}, g \in \widetilde{\mathcal{F}_n}$. We begin with the following inequality, which can be seen by splitting the expectation in the part where $\{f = g\}$ and $\{f \neq g\}$ and using the fact that $\operatorname{supp} \mathbb{P} \subset B(0,A)$:

$$\begin{aligned} |\mathrm{WSS}(f) - \mathrm{WSS}(g)| &= |\mathbb{E}\textstyle\sum_{k=1}^K f_k(X)\|X - c_k(f)\|^2 - g_k(X)\|X - c_k(g)\|^2| \\ &\leq 4A^2 d(f,g) + \textstyle\int_{\{f=g\}} \sum_{k=1}^K f_k(X)\left(\|X - c_k(f)\|^2 - \|X - c_k(g)\|^2\right). \end{aligned}$$

For the second term we have already seen in the proof of the previous lemma that $\|X - c_k(f)\|^2 - \|X - c_k(g)\|^2 \leq 6A\|c_k(f) - c_k(g)\|$. So for the moment we have

$$|\mathrm{WSS}(f) - \mathrm{WSS}(g)| \leq 4A^2 d(f,g) + 6A \sup_k \|c_k(f) - c_k(g)\|.$$

Now we want to bound the expression $\|c_k(f) - c_k(g)\|$. First of all, observe that $|\mathbb{E} f_k(X) - g_k(X)| \leq d(f,g)$ and $\|\mathbb{E} f_k(X)X - g_k(X)X\| \leq Ad(f,g)$.

In case $\mathbb{E} g_k(X) \neq 0$ we have

$$\begin{aligned} \|c_k(f) - c_k(g)\| &= \frac{\|\mathbb{E} g_k(X)\mathbb{E} f_k(X)X - \mathbb{E} f_k(X)\mathbb{E} g_k(X)X\|}{\mathbb{E} f_k(X)\mathbb{E} g_k(X)} \\ &\leq \frac{\|\mathbb{E} g_k(X)\left(\mathbb{E} f_k(X)X - \mathbb{E} g_k(X)X\right)\| + \|\left(\mathbb{E} g_k(X) - \mathbb{E} f_k(X)\right)\mathbb{E} g_k(X)X\|}{\mathbb{E} f_k(X)\mathbb{E} g_k(X)} \\ &\leq \frac{\mathbb{E} g_k(X)\|\mathbb{E} f_k(X)X - g_k(X)X\| + A\mathbb{E} g_k(X)|\mathbb{E} g_k(X) - f_k(X)|}{\mathbb{E} f_k(X)\mathbb{E} g_k(X)} \\ &\leq \frac{2A}{\mathbb{E} f_k(x)} d(f,g) \\ &\leq \frac{2A}{a} d(f,g). \end{aligned}$$

On the other hand, in case $\mathbb{E} g_k(X) = 0$ we also have $\mathbb{E} g_k(X)X = 0$ (as g_k is a non-negative function and $|X|$ is bounded by A). This leads to

$$\|c_k(f) - c_k(g)\| = \|\frac{\mathbb{E} f_k(X)X}{\mathbb{E} f_k(X)} - \frac{\mathbb{E} g_k(X)X}{\mathbb{E} g_k(X)}\| = \|\frac{\mathbb{E} f_k(X)X}{\mathbb{E} f_k(X)}\| \leq \frac{A}{a} d(f,g) \leq \frac{2A}{a} d(f,g).$$

Combining all results leads to

$$|\mathrm{WSS}(f) - \mathrm{WSS}(g)| \leq 4A^2(1 + 3/a) d(f,g)$$

which proves the lemma. ■

Lemma 22 (Condition (1) for Mod**)** *If* $m^2 \log n / n \to 0$, *then* Mod *satisfies Condition (1) of Theorem* 2.

Proof Let $f \in \widetilde{f}$. Using McDiarmid inequality one can prove

$$\mathbb{P}\left(\left|\sum_{k=1}^K \frac{1}{n(n-1)} \sum_{i \neq j} f_k(X_i) f_k(X_j) s(X_i, X_j) - \sum_{k=1}^K \mathbb{E} f_k(X) f_k(Y) s(X,Y)\right| \geq \varepsilon\right) \leq 2e^{-\frac{n\varepsilon^2}{2C^2K^2}}.$$

Now for ease of notation let

$$Q_n(f) = \frac{1}{n(n-1)^3}\sum_{k=1}^{K}\sum_{i\neq j} f_k(X_i)f_k(X_j)\sum_{l,l\neq i} s(X_i,X_l)\sum_{l,l\neq j} s(X_j,X_l),$$
$$\widetilde{Q_n}(f) = \frac{1}{n(n-1)}\sum_{k=1}^{K}\sum_{i\neq j} f_k(X_i)f_k(X_j)\int s(X_i,Z)d\mathbb{P}(Z)\int s(X_j,Z)d\mathbb{P}(Z),$$
$$Q(f) = \sum_{k=1}^{K}\int\int f_k(X)f_k(Y)\int s(X,Z)d\mathbb{P}(Z)\int s(Y,Z)d\mathbb{P}(Z)d(\mathbb{P}\times\mathbb{P})(X,Y).$$

If we have an exponential bound for $\mathbb{P}(|Q_n(f)-Q(f)|\geq\varepsilon)$ then with the above bound we would have an exponential bound for $\mathbb{P}(|\mathrm{Mod}_n(f)-\mathrm{Mod}(f)|\geq\varepsilon)$. Thus with the same argument than the one at the end of Lemma 14 the current lemma will be proved.
First note that

$$\mathbb{P}(|Q_n(f)-Q(f)|\geq\varepsilon)\leq\mathbb{P}(|Q_n(f)-\widetilde{Q_n}(f)|\geq\varepsilon/2)+\mathbb{P}(|\widetilde{Q_n}(f)-Q(f)|\geq\varepsilon/2).$$

Moreover $\mathbb{E}\widetilde{Q_n}(f)=Q(f)$ and thus with McDiarmid one can prove that

$$\mathbb{P}(|\widetilde{Q_n}(f)-Q(f)|\geq\varepsilon)\leq 2e^{\frac{n\varepsilon^2}{2C^4K^2}}.$$

The next step is to use the fact that for real numbers $a,b,a_n,b_n\in B(0,C)$,

$$|ab-a_nb_n| = |ab-a_nb+a_nb-a_nb_n|\leq C(|a-a_n|+|b-b_n|).$$

This implies the following inequalities:

$$\begin{aligned}
&|Q_n(f)-\widetilde{Q_n}(f)|\\
&\leq\frac{K}{n(n-1)}\sum_{i\neq j}\left|\frac{1}{(n-1)^2}\sum_{l,l\neq i}s(X_i,X_l)\sum_{l,l\neq j}s(X_j,X_l)-\int s(X_i,Z)d\mathbb{P}(Z)\int s(X_j,Z)d\mathbb{P}(Z)\right|\\
&\leq 2CK\sup_i\left|\frac{1}{n-1}\sum_{l,l\neq i}s(X_i,X_l)-\int s(X_i,Z)d\mathbb{P}(Z)\right|.
\end{aligned}$$

Hence the following:

$$\begin{aligned}
\mathbb{P}(|Q_n(f)-\widetilde{Q_n}(f)|\geq\varepsilon) &\leq \mathbb{P}(\sup_i|\frac{1}{n-1}\sum_{l,l\neq i}s(X_i,X_l)-\int s(X_i,Z)d\mathbb{P}(Z)|\geq\varepsilon/(2CK))\\
&\leq n\sup_i\mathbb{P}(|\frac{1}{n-1}\sum_{l,l\neq i}s(X_i,X_l)-\int s(X_i,Z)d\mathbb{P}(Z)|\geq\varepsilon/(2CK)).
\end{aligned}$$

Now to bound the last term we condition on X_i and use the McDiarmid inequality. Then taking the expectation yields the exponential bound:

$$\begin{aligned}
&\mathbb{P}(|\frac{1}{n-1}\sum_{l,l\neq i} s(X_i,X_l) - \int s(X_i,Z)d\mathbb{P}(Z)| \geq \varepsilon/(2CK)) \\
&= \mathbb{E}(\mathbb{P}(|\frac{1}{n-1}\sum_{l,l\neq i} s(X_i,X_l) - \int s(X_i,Z)d\mathbb{P}(Z)| \geq \varepsilon/(2CK)|X_i)) \\
&\leq \mathbb{E}(2e^{-\frac{n\varepsilon^2}{2C^4K^2}}) \\
&= 2e^{-\frac{n\varepsilon^2}{2C^4K^2}}.
\end{aligned}$$

All in all we proved that

$$\mathbb{P}(|\operatorname{Mod}_n(f) - \operatorname{Mod}(f)| \geq \varepsilon) \leq 2e^{-\frac{n\varepsilon^2}{8C^2K^2}} + 2(n+1)e^{-\frac{n\varepsilon^2}{32C^2K^2}}.$$

The n in front of the exponential obviously does not matter for the limit, see end of the proof of Lemma 14. ■

Lemma 23 (Condition (3) for Mod) Mod *satisfies Condition (3) of Theorem* 2.

Proof Let $f \in \mathcal{F}, g \in \widetilde{\mathcal{F}_n}$. Following the proof of Lemma 15 we have:

$$\begin{aligned}
|\operatorname{Mod}(f) - \operatorname{Mod}(g)| &\leq \sum_{k=1}^{K} \int\int_{(\{f=g\}^2)^c} (C+C^2) \\
&= K(C+C^2)(1-(1-d(f,g))^2) \\
&\leq 2K(C+C^2)d(f,g).
\end{aligned}$$

■

A.7 The Proofs of the Convergence Rates in Theorems 4 and 6

The following lemma collects all the bounds given in the previous proofs for WSS. Whenever possible, we used the one-sided McDiarmid inequality.

Lemma 24 *Assume that* $\operatorname{supp}\mathbb{P} \subset B(0,A)$ *for some constant* $A > 0$. *Let* $a_n^* := \inf_k \mathbb{E} f_k^*(X) - a_n$. *Then* $a_n^* \to a^* := \inf_k \mathbb{E} f_k^*(X) - a > 0$. *For all* n *and* $\varepsilon > 0$ *there exists a constant* $b(a_n^*/2)$ *which tends to a constant* $C' > 0$ *when* $n \to \infty$, *and a constant* $b(\varepsilon/(8A^2(1+3/a)))$ *(see Lemma* 11 *for more details about b) such that the following holds true*

$$\begin{aligned}
&\mathbb{P}(|\operatorname{WSS}(f_n) - \operatorname{WSS}(f^*)| \geq \varepsilon) \\
&\leq 2K^{m+1}(2n)^{(d+1)m^2}\left(\frac{4dKe^{-\frac{na^2\varepsilon}{616d\max(1,A^2)A^2(A+1)^2}} + 2e^{-\frac{n\varepsilon^2}{32A^4}}}{1-4dKe^{-\frac{na^2\varepsilon}{308d\max(1,A^2)A^2(A+1)^2}} - 2e^{-\frac{n\varepsilon^2}{8A^4}}} + \frac{Ke^{-\frac{n(a_n-a)^2}{8}}}{1-e^{-\frac{n(a_n-a)^2}{2}}} + \frac{Ke^{-\frac{na_n^{*2}}{32}}}{1-e^{-\frac{na_n^{*2}}{8}}}\right) \\
&+ \frac{4K}{a_n^*}e^{-mb(a_n^*/2)} + (16A^2(1+3/a)/\varepsilon)e^{-mb(\varepsilon/(8A^2(1+3/a)))}.
\end{aligned}$$

A.8 Proof of Theorem 4

First we take care of the last two terms. There exists N' which depends on the rate of convergence of a_n and on a^* such that for $n \geq N'$ we have

$$a_n^* \leq a^*/2.$$

This implies $b(a_n^*/2) \leq b(a^*/4)$ (see Lemma 11 for details). Now let $C_1' := b(\varepsilon/(8A^2(1+3/a)))$ and $C_2' := b(a^*/4)$. Then for $n \geq N'$ we have:

$$\begin{aligned}
&\frac{4K}{a_n^*} e^{-mb(a_n^*/2)} + (16A^2(1+3/a)/\varepsilon) e^{-mb(\varepsilon/(8A^2(1+3/a)))} \\
&\leq 8Ka^* e^{-C_2' m} + (16A^2(1+3/a)/\varepsilon) e^{-C_1' m} \\
&\leq C_1 e^{-C_2 m}
\end{aligned}$$

with

$$C_1 := \max(8Ka^*; 16A^2(1+3/a)/\varepsilon) \qquad \text{and} \qquad C_2 := \min(C_1'; C_2').$$

C_2 is a positive constant which depends on a, a^*, A, ε and $\mathbb{P}$. C_1 depends on K, a, a^*, ε and A.

Since we assume $n(a_n - a)^2 \to \infty$ there exists N'' which depends on the rate of convergence of a_n and on a^* such that $n \geq N''$ implies:

$$e^{-\frac{n(a_n-a)^2}{8}} \leq 1/2 \qquad \text{and} \qquad e^{-\frac{na_n^*}{32}} \leq e^{-\frac{n(a_n-a)^2}{8}}.$$

This means that for $n \geq N''$:

$$\frac{Ke^{-\frac{n(a_n-a)^2}{8}}}{1-e^{-\frac{n(a_n-a)^2}{2}}} + \frac{Ke^{-\frac{na_n^{*2}}{32}}}{1-e^{-\frac{na_n^{*2}}{8}}} \leq 4Ke^{-\frac{n(a_n-a)^2}{8}}.$$

Finally let $N = \max(N', N'')$ and

$$\begin{aligned}
C_3 &:= \frac{8dK}{1 - 4dKe^{-\frac{Na^2\varepsilon}{308d\max(1,A^2)A^2(A+1)^2}} - 2e^{-\frac{N\varepsilon^2}{8A^4}}} \\
C_4 &:= \min\left(\frac{a^2}{616d\max(1,A^2)A^2(A+1)^2}; \frac{1}{32A^4}\right).
\end{aligned}$$

Since $\varepsilon \leq 1$ we have with these notations for $n \geq N$:

$$\frac{4dKe^{-\frac{na^2\varepsilon}{616d\max(1,A^2)A^2(A+1)^2}} + 2e^{-\frac{n\varepsilon^2}{32A^4}}}{1 - 4dKe^{-\frac{na^2\varepsilon}{308d\max(1,A^2)A^2(A+1)^2}} - 2e^{-\frac{n\varepsilon^2}{8A^4}}} \leq (C_3/2)e^{-C_4\varepsilon^2 n}.$$

All in all Theorem 4 is proved. ■

The **Proof of Theorem 6** works analogously, we just replace the above lemma by to following one:

Lemma 25 *Assume that the similarity function s is bounded by $C > 0$. Let $a_n^* := \inf_k \mathrm{vol}(f_k^*) - a_n$. Then $a_n^* \to \inf_k \mathrm{vol}(f_k^*) - a > 0$. For all n and $\varepsilon > 0$ there exists a constant $b(a_n^*/(2S))$ which tends to a constant $C' > 0$ when $n \to \infty$, and a constant $b(a\varepsilon/(8SK))$ (see Lemma 11 for more details about b) such that the following holds true*

$$\mathbb{P}(|\mathrm{Ncut}(f_n) - \mathrm{Ncut}(f^*)| \geq \varepsilon)$$

$$\leq 2K^{m+1}(2n)^{(d+1)m^2} \left(\frac{4e^{-\frac{na^2\varepsilon^2}{2048C^2K^2}}}{1-4Ke^{-\frac{na^2\varepsilon^2}{512C^2K^2}}} + \frac{e^{-\frac{n(a_n-a)^2}{32C^2}}}{1-e^{-\frac{n(a_n-a)^2}{8C^2}}} + \frac{e^{-\frac{na_n^{*2}}{128C^2}}}{1-e^{-\frac{na_n^{*2}}{32C^2}}} \right)$$

$$+ \frac{4CK}{a_n^*} e^{-mb(a_n^*/(2C))} + \frac{16CK}{a\varepsilon} e^{-mb(a\varepsilon/(8CK))}.$$

References

S. Ben-David. A framework for statistical clustering with constant time approximation algorithms for k-median and k-means clustering. *Machine Learning*, 66:243 – 257, 2007.

J. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical report, University of Bonn, 1998.

A. Czumaj and C. Sohler. Sublinear-time approximation algorithms for clustering via random sampling. *Random Struct. Algorithms*, 30(1-2):226–256, 2007.

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

M. Figueiredo and A. Jain. Unsupervised learning of finite mixture models. *PAMI*, 24(3):381–396, 2002.

C. Fraley and A. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J*, 41(8):578–588, 1998.

J. Fritz. Distribution-free exponential error bound for nearest neighbor pattern classification. *IEEE Trans. Inf. Th.*, 21(5):552 – 557, 1975.

M. Garey, D. Johnson, and H. Witsenhausen. The complexity of the generalized Lloyd - max problem (corresp.). *IEEE Trans. Inf. Theory*, 28(2):255–256, 1982.

P. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

S. Guattery and G. Miller. On the quality of spectral separators. *SIAM Journal of Matrix Anal. Appl.*, 19(3):701 – 719, 1998.

J. Hartigan. Consistency of single linkage for high-density clusters. *JASA*, 76(374):388 – 394, 1981.

J. Hartigan. Statistical theory in clustering. *Journal of Classification*, 2:63 – 76, 1985.

M. Inaba, N. Katoh, and H. Imai. Applications of weighted Voronoi diagrams and randomization to variance-based *k*-clustering. In *Proceedings of the 10th Annual Symposium on Computational Geometry*, pages 332–339. ACM Press, Stony Brook, USA, 1994.

P. Indyk. Sublinear time algorithms for metric space problems. In *Proceedings of the Thirty-first Annual ACM Symposium on Theory of Computing (STOC)*, pages 428–434. ACM Press, New York, 1999.

S. Jegelka. Statistical learning theory approaches to clustering. Master's thesis, University of Tübingen, 2007. Available at http://www.kyb.mpg.de/publication.html?user=jegelka.

R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, 2004.

C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148 – 188, 1989. Cambridge University Press.

G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley, New York, 2004.

N. Mishra, D. Oblinger, and L. Pitt. Sublinear time approximate clustering. In *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA-01)*, pages 439–447. ACM Press, New York, 2001.

M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74:036104, 2006.

D. Pollard. Strong consistency of k-means clustering. *Annals of Statistics*, 9(1):135 – 140, 1981.

A. Rakhlin and A. Caponnetto. Stability of *k*-means clustering. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

D. Spielman and S. Teng. Spectral partitioning works: planar graphs and finite element meshes. In *37th Annual Symposium on Foundations of Computer Science (Burlington, VT, 1996)*, pages 96 – 105. IEEE Comput. Soc. Press, Los Alamitos, CA, 1996. (See also extended technical report version.).

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395 – 416, 2007.

U. von Luxburg and S. Ben-David. Towards a statistical theory of clustering. In *PASCAL workshop on Statistics and Optimization of Clustering, London*, 2005.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *Annals of Statistics*, 36(2):555 – 586, 2008.

U. von Luxburg, S. Bubeck, S. Jegelka, and M. Kaufmann. Consistent minimization of clustering objective functions. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems (NIPS) 21*. MIT Press, Cambridge, MA, 2008.

D. Wagner and F. Wagner. Between min cut and graph bisection. In *Proceedings of the 18th International Symposium on Mathematical Foundations of Computer Science (MFCS)*, pages 744 – 750, London, 1993. Springer.

M. Wong and T. Lane. A kth nearest neighbor clustering procedure. *J.R. Statist.Soc B*, 45(3): 362 – 368, 1983.