Supplementary material to "Consistent Minimization of Clustering Objective Functions"

Ulrike von Luxburg
Max Planck Institute for Biological Cybernetics
ulrike.luxburg@tuebingen.mpg.de

Sébastien Bubeck
INRIA Futurs Lille, France
sebastien.bubeck@inria.fr

Stefanie Jegelka
Max Planck Institute for Biological Cybernetics
stefanie.jegelka@tuebingen.mpg.de

Michael Kaufmann
University of Tübingen, Germany
mk@informatik.uni-tuebingen.de

Abstract

This paper contains supplementary material to the paper "Consistent Minimization of Clustering Objective Functions", published at NIPS 2007 (von Luxburg et al., to appear). In particular, this supplement contains detailed proofs of all the theorems. For the general background please see von Luxburg et al. (to appear). An extended version which combines most of the NIPS paper and this supplement, and contains more general results and a more elegant approach to the proofs is available as a preprint (Bubeck and von Luxburg, 2007). For readers who are interested in our results in general we suggest to read the preprint (Bubeck and von Luxburg, 2007) rather than this supplement. Readers who want to see the proofs of the theorems exactly as they have been formulated in the NIPS paper (von Luxburg et al., to appear) should read this supplement.

1 General setup and notation

Intuitively, a clustering should discover "meaningful" groups. As we think that those groups should be "connected" in some sense, we will only consider clusterings which are continuous in a certain sense. More formally, we represent a clustering of $\mathcal{X}$ by a function from $\mathcal{X}$ to $\{1, \ldots, K\}$ which is almost surely (with respect to $\mathbb{P}$) continuous. We will always fix the number K of clusters (and we will not touch the question of what "the best K" is). Given a certain clustering quality function Q, the goal of clustering is to find a function which minimizes Q. To be as general as possible we also allow for the possibility that one has additional constraints on the clustering solution. For example, on could imagine to only look for clusters which have a certain "minimal size". We will express such additional constraints by introducing a "predicate" $A(f)$, that is A denotes a property of the function which can be either "true" or "false". The "true clustering" f^* is then defined as a function f which has minimal value $Q(f)$ among all functions which are continuous almost everywhere and satisfy $A(f)$. Given a finite sample, we now want to approximate this true clustering. As we do not know $\mathbb{P}$ in this setting, we can neither evaluate Q nor A. Instead, we introduce approximations $Q_n(f)$ and $A_n(f)$ of the two quantities. Moreover, according to the discussion above we need to restrict the space of functions to some space $\mathcal{F}_n$. The empirical clustering f_n is then defined as the function which has minimal value $Q_n(f)$ among all functions which are in $\mathcal{F}_n$ and satisfy $A_n(f)$.

In the following we always assume that the underlying space is $\mathcal{X} = \mathbb{R}^d$, endowed with the Euclidean distance. We will use the following notations and abbreviations :

d	dimension of the space $\mathbb{R}^d$;
n	number of sample points;
$X_1, \ldots, X_n$	sample points drawn i.i.d from $\mathbb{P}$;
$m := m(n) \leq n$	number of seeds used in the nearest neighbor operation. This number $m(n)$ will depend on n, but for readability reasons we will drop the argument in the following and simply write m instead;
K	number of clusters to construct;
$NN_m(x)$	is the nearest neighbor of x among $X_1, \ldots, X_n$ according to the euclidean distance;
$B(x, \delta)$	the ball around x with radius δ
$\mathcal{H}$	is the space of all functions from $\mathbb{R}^d$ to $\{1, \ldots, K\}$;
$A : \mathcal{H} \to \{True, False\}$	is a predicate, that is a certain property of a function which can be true or false. If f is a data dependent function then we define $\mathbb{P}(A(f)) := \mathbb{P}(\{A(f) = True\})$ and $\mathbb{P}(A(f)^c) := \mathbb{P}(\{A(f) = False\})$
$A_n : \mathcal{H} \times (\mathbb{R}^d)^n \to \{True, False\}$	is a predicate depending on the function and the data points. Through the paper we will only be interested in $A_n(f, X_1, \ldots, X_n)$ and we will write it $A_n(f)$. We define $\mathbb{P}(A_n(f)) := \mathbb{P}(\{A_n(f) = True\})$ and $\mathbb{P}(A_n(f)^c) := \mathbb{P}(\{A_n(f) = False\})$.
$Q : \mathcal{H} \to \mathbb{R}^+$	the true clustering quality function (usually depends on $\mathbb{P}$)
$Q_n : \mathcal{H} \to \mathbb{R}^+$	the empirical clustering quality function which can be computed on the finite sample

For given properties $A(f)$ and $A_n(f)$ we will consider the following function spaces :

$$\mathcal{F} := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f \text{ continuous almost everywhere, } f \text{ satisfies } A(f)\}$$
$$\mathcal{F}_n := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid f \text{ satisfies } f(x) = f(NN_m(x)) \text{ and } A_n(f)\}$$
$$\widetilde{\mathcal{F}}_n := \bigcup_{X_1, \ldots, X_n \in \mathbb{R}^d} \mathcal{F}_n$$
$$\widehat{\mathcal{F}}_n := \{f : \mathbb{R}^d \to \{1, \ldots, K\} \mid \exists \text{ Voronoi partition of } m \text{ cells: } f \text{ constant on all cells}\}$$

We will endow all function spaces with the pseudo-distance

$$d(f, g) := \mathbb{P}(f(X) \neq g(X) \mid X_1, \ldots, X_n)$$

Note that we need the conditioning in case f or g are data dependent. The conditioning has no effect if this is not the case.

Given all the quantities mentioned above, we define the true and the empirical clusterings as

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

Note that we do not assume that f_n or f^* are unique. Lastly we define the functions :

$$f_n^* \in \operatorname*{argmin}_{f \in \mathcal{F}_n} Q(f)$$
$$\widetilde{f}^*(x) = f^*(NN_m(x))$$

Note that to avoid technical overload we will assume throughout this paper that all the suprema and infima are attained. If this is not the case one can always go over to statements about functions which are ε-close to the suprema or infima, respectively. However, this leads to a heavy overload in notation, which we would like to avoid. The reader will be able to modify our results appropriately if necessary.

Morevoer, it is very important to point out that the functions $f_n, \tilde{f}^*$ and the function space $\mathcal{F}_n$ are data-dependent. Thus we have to be very careful when dealing with probabilities related to those quantities.

2 Main result: consistency theorem in a general setting

The following theorem is the main theorem of our paper. It states general assumptions on the quantities involved which ensure that nearest neighbor clustering is weakly consistent.

Theorem 1 (Consistency of nearest neighbor clustering) *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 $\mathbb{R}^d$ (with $d > 1$). Let $m := m(n) \leq n$ be the number of points used as nearest neighbor centers in the nearest neighbor clustering algorithm. For any function $f \in \mathcal{H}$ let $A(f)$ and $A_n(f)$ be predicates as defined above. 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 (that is, it does not involve any function evaluations $f(x)$ for $x \notin \{X_1, \ldots, X_n\}$). Assume that the following conditions are satisfied:*

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

$$\forall \epsilon > 0, \quad K^m (2n)^{(d+1)m^2} \sup_{f \in \widetilde{\mathcal{F}}_n} \mathbb{P}(|Q_n(f) - Q(f)| > \epsilon) \to 0,$$

2. *$A_n(f)$ is a consistent estimator of $A(f)$ in the following sense :*

$$\mathbb{P}(A_n(\widetilde{f^*})) \to 1$$
$$\mathbb{P}(A(f_n)) \to 1,$$

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

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

4. *$lim_{n\to\infty} m(n) = +\infty$.*

Then nearest neighbor clustering is weakly consistent, that is $Q(f_n)$ tends to $Q(f^*)$ in probability.

3 Proof of Theorem 1

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

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

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

$$\mathbb{P}(Q(f_n) - Q(f^*) \leq -\epsilon) \leq \mathbb{P}(f_n \notin \mathcal{F}) = \mathbb{P}(A(f_n)^c).$$

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 $\epsilon > 0$ we have :

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

In the following sections we will treat both parts separately. In the remaining part of this section we will always assume that the "general assumptions" of Theorem 1 are true (that is the assumptions except the ones labeled by (1)-(4)).

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

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

Now we will focus on the right hand side of this inequality. In order to bound it we will state two lemmas. The proof techniques for those lemmas are similar to the ones in Devroye et al. (1996), Section 12.3 and Chapter 13. The first lemma is the most important part of the proof. It allows us to get the supremum out of the probability. Recall that the shattering coefficient of a function class $\mathcal{G}$ of functions in a discrete valued space $\{1, ..., K\}$ with respect to sample size n is defined by

$$s(\mathcal{G}, n) = \max_{x_1, \ldots, x_n} |\{(f(x_1), \ldots, f(x_n)) \mid f \in \mathcal{G}\}|.$$

Lemma 2 *Under the general assumptions of Theorem* 1 *we have:*

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

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

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

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 $|Q_n(\widehat{f}) - Q(\widehat{f})| \geq \epsilon$; if such an $\widehat{f}$ does not exist then just choose $\widehat{f}$ as some other fixed function in $\widetilde{\mathcal{F}}_n$. Notice that $\widehat{f}$ is a data-dependent function depending on the sample $X_1, ..., X_n$.

We have the following inequalities:

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

The last step is true because of the definition of $\hat{f}$ (note that due to the definition of $\hat{f}$ the event $|Q_n(\hat{f}) - Q(\hat{f})| \geq \varepsilon$ is true iff there exists some $f \in \tilde{\mathcal{F}}_n$ such that $|Q_n(f) - Q(f)| \geq \varepsilon$, which is true iff

$\sup_{f\in\tilde{\mathcal{F}}_n} |Q_n(f) - Q(f)| \geq \varepsilon)$.

Rearranging the inequality of above leads to

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

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:

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

This completes the proof of the lemma. □

Remark *Readers familiar with symmetrization might wonder where the unusual term involving the infimum in the denominator comes from. In fact, it pops up 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.*

Lemma 3 *Let $u \in \mathbb{N}$ and $\widetilde{\mathcal{F}}_n$ and $\widehat{\mathcal{F}}_n$ be the function sets defined above. 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 $d > 1$.

□

Combining the lemmas above now leads to a bound on the estimation error:

Proposition 4 (Bound on the estimation error) *Under the general assumptions of Theorem 1 we have for every fixed $\varepsilon > 0$ that*

$$\mathbb{P}(Q(f_n) - Q(f_n^*) \geq \epsilon) \leq 2K^m(2n)^{(d+1)m^2} \frac{\sup_{f\in\widetilde{\mathcal{F}}_n} \mathbb{P}(|Q_n(f) - Q(f)| \geq \epsilon/8)}{\inf_{f\in\widetilde{\mathcal{F}}_n} \mathbb{P}(|Q_n(f) - Q(f)| \leq \epsilon/4)}.$$

Under Assumption (1) of Theorem 1 this converges to 0.

Proof. With the comment at the beginning of this section we have

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

Then the results of Lemmas 2 and 3 yields the bound. 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. □

3.2 Approximation error

Recall that $\widetilde{f}^*$ is defined by $\widetilde{f}^*(x) = f^*(NN_m(x))$. If $A_n(\widetilde{f}^*)$ is true then $\widetilde{f}^* \in \mathcal{F}_n$. By the definition of f_n^* then the following bound holds true :

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

This leads to the following:

$$\begin{aligned}\mathbb{P}(Q(f_n^*) - Q(f^*) \geq \epsilon) &= \mathbb{P}(Q(f_n^*) - Q(f^*) \geq \epsilon, f^* \in \mathcal{F}_n) + \mathbb{P}(Q(f_n^*) - Q(f^*) \geq \epsilon, f^* \notin \mathcal{F}_n) \\ &\leq \mathbb{P}(\widetilde{f}^* \in \mathcal{F}_n, Q(\widetilde{f}^*) - Q(f^*) \geq \epsilon) + \mathbb{P}(A_n(\widetilde{f}^*)^c)\end{aligned}$$

The second term converges to 0 due to Assumption (2) of Theorem 1. To bound the first term we will now use Assumption (3) of Theorem 1. We first prove the following lemma with ideas from Fritz, 1975.

Lemma 5 *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 $\epsilon > 0$ there exists a constant $b(\epsilon) > 0$ independent of n and f such that

$$\mathbb{P}(L_n \geq \epsilon) \leq \frac{2}{\epsilon} e^{-mb(\epsilon)}.$$

Proof. The first step of the proof consists in constructing a certain set B (depending on ε) which satisfies the following statement:

For all ε there exists a $\delta(\varepsilon) > 0$, a measurable set $B \subset \mathbb{R}^d$ and a constant $1 > u > 0$ such that
(a) $\mathbb{P}(B) \geq 1 - \varepsilon/2$
(b) $\forall x \in B : \mathbb{P}(B(x, \delta)) > u$
(c) $\forall x \in B$ the function f is constant on $B(x, \delta)$.

Assume we have such a set B. Then using Property (c) and (a) we see that

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

Using the Markov inequality we can then see that

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

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

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

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

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

So to finish the proof of the lemma we have to show how the set B 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 $\mathbb{P}(A_{\delta(\varepsilon)}) \geq 1 - \epsilon/4$. 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 $\mathbb{P}(U) = 0$ which proves our statement.

In other words we have $\mathbb{P}(\mathbb{P}(B(X, \delta)|X) > 0) = 1$, 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 $\mathbb{P}(A) \geq 1 - \epsilon/4$ and $\forall x \in A, \mathbb{P}(B(x, \delta)) > u$. So all points in A satisfy Property (b).

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

Proposition 6 (Bound on the approximation error) *If the general assumptions of Theorem 1 and Assumption* (3) *are satisfied, then for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ and a $b(\delta(\varepsilon)) > 0$ such that*

$$\mathbb{P}(Q(f_n^*) - Q(f^*) \geq \epsilon) \leq \mathbb{P}(A_n(\widetilde{f^*})^c) + \frac{2}{\delta(\epsilon)} e^{-mb(\delta(\epsilon))}.$$

Moreover, under Assumptions (2) *and* (4) *this term tends to* 0 *if $n \to \infty$ and $\varepsilon > 0$ is fixed.*

Proof. With the comment made at the beginning of this section we have

$$\mathbb{P}(Q(f_n^*) - Q(f^*) \geq \epsilon) \leq \mathbb{P}(A_n(\widetilde{f^*})^c) + \mathbb{P}(\widetilde{f^*} \in \mathcal{F}_n, Q(\widetilde{f^*}) - Q(f^*) \geq \epsilon).$$

Now if $\widetilde{f^*} \in \mathcal{F}_n$ then by Assumption (3) we know that $\forall \epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that

$$d(\widetilde{f^*}, f^*) \leq \delta(\varepsilon) \Rightarrow Q(\widetilde{f^*}) - Q(f^*) \leq \epsilon.$$

So with the previous lemma it means that :

$$\mathbb{P}(\widetilde{f^*} \in \mathcal{F}_n, Q(\widetilde{f^*}) - Q(f^*) \geq \epsilon) \leq \mathbb{P}\left(\mathbb{P}(d(\widetilde{f^*}, f^*) > \delta(\epsilon))\right) \leq \frac{2}{\delta(\epsilon)} e^{-mb(\delta(\epsilon))}.$$

□

3.3 Rate of convergence

Actually, the proof of Theorem 1 implies the following bound:

Theorem 7 (Rate of Convergence) *Under the general assumptions of Theorem 1 and if Assumption* (3) *is satisfied then for all $\varepsilon > 0$ there exist constants $\delta(\varepsilon) > 0$ and $b(\delta(\varepsilon)) > 0$ such that for all $n \in \mathbb{N}$ and $m \leq n$ we have*

$$\mathbb{P}(|Q(f_n) - Q(f^*)| \geq \epsilon)$$

$$\leq 2K^m (2n)^{(d+1)m^2} \frac{\sup_{f \in \widetilde{\mathcal{F}}_n} \mathbb{P}(|Q_n(f) - Q(f)| \geq \epsilon/16)}{\inf_{f \in \widetilde{\mathcal{F}}_n} \mathbb{P}(|Q_n(f) - Q(f)| \leq \epsilon/8)} + \mathbb{P}(A_n(\widetilde{f^*})^c) + \mathbb{P}(A(f_n)^c) + \frac{2}{\delta(\epsilon/2)} e^{-mb(\delta(\epsilon/2))}.$$

Remark *Note that the constant $\delta(\varepsilon)$ is the one given in Assumption (3) of Theorem 1, whereas the constant $b(\delta(\varepsilon))$ has been defined only implicitly in Lemma 5. As we do not have an explicit functional form of the dependency of b on ε we cannot simply use the Borel-Cantelli lemma to transform the preceding bound into almost sure convergence. If one would like to prove strong consistency of nearest neighbor clustering one would have to get an explicit form of $\delta(\varepsilon)$ and $b(\delta(\varepsilon))$. Of course, to apply Borel-Cantelli one would also need rates of convergence in Assumption (2) of Theorem 1.*

4 Applications

Now we will apply our theorem to different popular clustering objective functions and show that nearest neighbor clustering is consistent under very mild conditions. The major part of the work is often to prove that Assumption (1) of Theorem 1 is true. In cases where $\mathbb{E}Q_n = Q$ this can be very simple by using concentration inequalities. However, if $\mathbb{E}Q_n \neq Q$ we have to use more elaborate techniques. In fact, it turns out that this is the case for all the clustering objective functions we are going to study. The second issue is proving Assumption (2) of Theorem 1. In our examples we will focus on properties $A(f)$ which state that the clustering function f constructs K clusters of a certain "minimal size". In the following we would like to introduce some general propositions to deal with this case. Again we need some notation:

Introduce operators $\Phi_k, \Phi_{k,n} : \mathcal{H} \to \mathbb{R}^d$. Let $a > 0$ be a fixed constant and $(a_n)_{n\in\mathbb{N}}$ a sequence with $a_n > a$ and $a_n \to a$. For any $g \in \mathcal{H}$ let

$$A_n(g) = true \iff \forall k \in \{1, \dots, K\}, \Phi_{k,n}(g) > a_n$$
$$A(g) = true \iff \forall k \in \{1, \dots, K\}, \Phi_k(g) > a.$$

In particular we have the following simple lemma:

Lemma 8 *Let $a > 0$, $a_n > a$, $a_n \to a$, and Φ_k be as defined above. Define*

$$a^* := \inf_k \Phi_k(f_k^*) - a$$
$$a_n^* := \inf_k \Phi_k(f_k^*) - a_n.$$

Then we have that $a^* > 0$. Furthermore, there exists some $N \in \mathbb{N}$ such that for all $n \geq N$ we have $a_n^* > 0$.

Proof. As $f^* \in \mathcal{F}$, the statement $a^* > 0$ follows from the definition of $A(f)$ (note that this the definition of $A(f)$ uses a strict inequality). The statement about a_n^* follows from the fact that $a_n \to a$ implies $a_n^* \to a^* > 0$. □

Proposition 9 (Assumption (2) of Theorem 1) *Assume that $\Phi_{k,n}$ can be computed on the data and that the following statements are true:*

(i) Φ_k is continuous in the following sense :

$$\forall \epsilon > 0, \exists \delta(\epsilon) : \forall f, g : \mathbb{R}^d \to \{1, \dots, K\}, d(f, g) \leq \delta \Rightarrow \Phi_k(f) - \Phi_k(g) \leq \epsilon,$$

(ii) $\Phi_{k,n}$ converges to Φ_k fast enough: $\forall k \in \{1, \dots, K\}, \forall \epsilon > 0$

$$K^m (2n)^{(d+1)m^2} \sup_{g \in \widehat{\mathcal{F}}_n} \mathbb{P}(\Phi_{k,n}(g) - \Phi_k(g) \geq \epsilon) \to 0,$$

(iii) a_n decreases slowly enough to a:

$$K^m (2n)^{(d+1)m^2} \sup_{g \in \mathcal{F}_n, k} \mathbb{P}(\Phi_{k,n}(g) - \Phi_k(g) \geq a_n - a) \to 0.$$

Then Assumption (2) of Theorem 1 is satisfied if $m \to \infty$.

Proof. We begin by proving that $\mathbb{P}(A(f_n)^c) \to 0$. As $f_n \in \mathcal{F}_n$ by definition we have that $\Phi_{k,n}(f_n) > a_n$ for all $k = 1, \dots, K$. A union bound argument shows that

$$\mathbb{P}(A(f_n)^c) \leq K \sup_k \mathbb{P}(\Phi_k(f_n) \leq a).$$

Using the same techniques as in the proof of Lemma 2 we can show:

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

Moreover, we already proved in Lemma 3 that $s(\widehat{\mathcal{F}}_n, 2n) \leq K^m (2n)^{(d+1)m^2}$ then under the Assumption (iii) we see that this term tends to 0.

Now we have to prove that $\mathbb{P}(A_n(\widetilde{f^*})^c) \to 0$. By the same union bound argument as above we have that

$$\mathbb{P}(A_n(\tilde{f}^*)^c) \leq K \sup_k \mathbb{P}(\Phi_{k,n}(\tilde{f}^*) \leq a_n).$$

According to Lemma 8 there exists N such that $n \geq N \Rightarrow a_n^* > 0$. Now for $n \geq N$ we have the following inequalities (using Assumption (i) of the proposition):

$$\begin{aligned}
&\mathbb{P}(A_n(\tilde{f}^*)^c) = \mathbb{P}(\widetilde{f^*} \notin \mathcal{F}_n) \\
&= \mathbb{P}(\Phi_{k,n}(\widetilde{f^*}) \leq a_n) \\
&= \mathbb{P}(\Phi_k(f^*) - \Phi_{k,n}(\widetilde{f^*}) \geq \Phi_k(f^*) - a_n) \\
&= \mathbb{P}(\Phi_k(f^*) - \Phi_k(\widetilde{f^*}) + \Phi_k(\widetilde{f^*}) - \Phi_{k,n}(\widetilde{f^*}) \geq \Phi_k(f^*) - a_n) \\
&\leq \mathbb{P}(\Phi_k(f^*) - \Phi_k(\widetilde{f^*}) \geq (\Phi_k(f^*) - a_n)/2) + \mathbb{P}(\Phi_k(\widetilde{f^*}) - \Phi_{k,n}(\widetilde{f^*}) \geq (\Phi_k(f^*) - a_n)/2) \\
&\leq \mathbb{P}(\Phi_k(f^*) - \Phi_k(\widetilde{f^*}) \geq a_n^*/2) + \mathbb{P}(\Phi_k(\widetilde{f^*}) - \Phi_{k,n}(\widetilde{f^*}) \geq a_n^*/2) \\
&\leq \mathbb{P}(d(f^*, \widetilde{f^*}) > \delta(a_n^*/2)) + \mathbb{P}(\sup_{g \in \widehat{\mathcal{F}}_n} \Phi_k(g) - \Phi_{k,n}(g) \geq a_n^*/2) \\
&\leq \tfrac{2}{\delta(a_n^*/2)} e^{-mb(\delta(a_n^*/2))} + \mathbb{P}(\sup_{g \in \widehat{\mathcal{F}}_n} \Phi_k(g) - \Phi_{k,n}(g) \geq a_n^*/2).
\end{aligned}$$

If $m \to \infty$ then the first term goes to 0. For the second term the key step is to see that doing exactly the same things that in proof of lemma 2 we get :

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

Under Assumption (ii) this term tends to 0. □

In the previous proof we implicitly also showed the following rate of convergence :

Proposition 10 (Rate of convergence for Assumption (2) of Theorem 1) *Assume that $\Phi_{k,n}$ can be computed on the data. Then we have*

$$\mathbb{P}(A(f_n)^c) \leq 2K^{m+1} (2n)^{(d+1)m^2} \sup_k \frac{\sup_{g \in \widehat{\mathcal{F}}_n} \mathbb{P}(\Phi_{k,n}(g) - \Phi_k(g) \geq (a_n - a)/4)}{\inf_{g \in \widehat{\mathcal{F}}_n} \mathbb{P}(\Phi_{k,n}(g) - \Phi_k(g) \leq (a_n - a)/2)}.$$

Moreover if Assumption (i) of Proposition 9 is satisfied and $n > N$ (recall the definitions of N and a_n^* in Lemma 8) then we have

$$\mathbb{P}(A_n(\widetilde{f^*})^c) \leq \frac{2K}{\delta(a_n^*/2)} e^{-mb(\delta(a_n^*/2))} + 2K^{m+1} (2n)^{(d+1)m^2} \sup_k \frac{\sup_{g \in \widehat{\mathcal{F}}_n} \mathbb{P}(\Phi_{k,n}(g) - \Phi_k(g) \geq a_n^*/8)}{\inf_{g \in \widehat{\mathcal{F}}_n, k} \mathbb{P}(\Phi_{k,n}(g) - \Phi_k(g) \leq a_n^*/4)}.$$

For convenience we also recall the McDiarmid inequality, which will be used several times.

Theorem 11 (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

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

4.1 Normalized cut objective function

In this section we want to apply the consistency theorem for nearest neighbor clustering to the case where the objective function is the normalized cut objective function. This is the objective function which spectral clustering attempts to minimize. In this section we assume that we are given a similarity function $s : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ which is upper bounded by a global constant $S > 0$. We make this assumption for convenience, as this will allow us at several places to apply the McDiarmid concentration inequality. Moreover we will use the convention "$0/0 = 0$" to define our objective functions. We define the $Ncut$ objective function as follows. For an indicator function $f : \mathbb{R}^d \to \{0, 1\}$ let

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

$$cut_n(f) := \frac{1}{n(n-1)}\sum_{i\neq j} f(X_i)(1 - f(X_j))s(X_i, X_j)$$

$$vol(f) := \mathbb{E}f(X_1)s(X_1, X_2)$$

$$vol_n(f) := \frac{1}{n(n-1)}\sum_{i\neq j} f(X_i)s(X_i, X_j)$$

For a clustering function $f : \mathbb{R}^d \to \{1, \ldots, K\}$ taking K values we define

$$f_k := 1_{f=k}$$

$$Ncut_n(f) := \sum_{k=1}^{K} \frac{cut_n(f_k)}{vol_n(f_k)}$$

$$Ncut(f) := \sum_{k=1}^{K} \frac{cut(f_k)}{vol(f_k)}.$$

Remark *Note that $\mathbb{E}Ncut_n \neq Ncut$ but $\mathbb{E}vol_n = vol$ and $\mathbb{E}cut_n = cut$.*

We consider the properties

$$A_n(f) = true \iff \forall k \in \{1, \ldots, K\}, vol_n(f_k) > a_n$$
$$A(f) = true \iff \forall k \in \{1, \ldots, K\}, vol(f_k) > a.$$

The reason why we need those properties is because the $Ncut$ objective function has the volume of the clusters in the denominator. The conditions now ensure that $Ncut$ is continuous on $\mathcal{F}_n$ and $\mathcal{F}$ (and even Lipschitz, so giving bounds will be easier). Note also that implicitly this conditions also implies that we always obtain exactly K clusters and do not allow empty clusters. For consistency reasons we will later assume that $a_n > a$ and $a_n \to a$.

To prove the consistency of nearest neighbor clustering for Ncut we now need to check the assumptions of Theorem 1. Clearly the general assumptions are true. Now we will prove one lemma for each for the remaining Assumptions (1) - (3).

Lemma 12 (Assumption 1 for Ncut) *For $f \in \widetilde{\mathcal{F}}_n$ we have*

$$\mathbb{P}(|Ncut(f) - Ncut_n(f)| > \epsilon) \leq 4Ke^{-\frac{na^2\epsilon^2}{8S^2K^2}}.$$

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

$$\{|cut_n(f_k) - cut(f_k)| \leq \tfrac{a}{2}\epsilon\} \bigcap \{|vol_n(f_k) - vol(f_k)| \leq \tfrac{a}{2}\epsilon\}$$

$$\subset \{|\tfrac{cut_n(f_k)}{vol_n(f_k)} - \tfrac{cut(f_k)}{vol(f_k)}| \leq \epsilon\}.$$

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

$$\begin{aligned}
\tfrac{cut_n(f_k)}{vol_n(f_k)} - \tfrac{cut(f_k)}{vol(f_k)} &= \tfrac{cut_n(f_k)vol(f_k) - cut(f_k)vol_n(f_k)}{vol_n(f_k)vol(f_k)} \\
&\leq \tfrac{(cut(f_k)+\epsilon)vol(f_k) - cut(f_k)(vol(f_k)-\epsilon)}{vol_n(f_k)vol(f_k)} \\
&= \tfrac{\epsilon}{vol_n(f_k)} \tfrac{cut(f_k)+vol(f_k)}{vol(f_k)} \\
&\leq \tfrac{2\epsilon}{a}.
\end{aligned}$$

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

$$\frac{cut_n(f)}{vol_n(f)} - \frac{cut(f)}{vol(f)} = \frac{cut_n(f)}{vol_n(f)} \leq \frac{\epsilon}{a} \leq \frac{2\epsilon}{a}.$$

Using the same technique we have the same bound for $\frac{cut(f_k)}{vol(f_k)} - \frac{cut_n(f_k)}{vol_n(f_k)}$, 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 $cut_n(f)$ and $vol_n(f)$ will change at most by $2S/n$. Together all this leads to

$$\begin{aligned}
&\mathbb{P}(|Ncut(f) - Ncut_n(f)| > \epsilon) \\
&\leq K \sup_k \mathbb{P}(|\frac{cut_n(f_k)}{vol_n(f_k)} - \frac{cut(f_k)}{vol(f_k)}| > \epsilon/K) \\
&\leq K \sup_k \Big(\mathbb{P}(|cut_n(f_k) - cut(f_k)| > \frac{a}{2K}\epsilon) + \mathbb{P}(|vol_n(f_k) - vol(f_k)| > \frac{a}{2K}\epsilon)\Big) \\
&\leq 4Ke^{-\frac{na^2\epsilon^2}{8S^2K^2}}.
\end{aligned}$$

□

Lemma 13 (Assumption (2) for Ncut) *The following holds true*

$$\mathbb{P}(A(f_n)^c) \leq 2K^{m+1}(2n)^{(d+1)m^2} \frac{e^{-\frac{n(a_n-a)^2}{32S^2}}}{1 - e^{-\frac{n(a_n-a)^2}{8S^2}}}.$$

If additionally we have $a_n^* > 0$ (recall the definition of a_n^* in Lemma 8) then

$$\mathbb{P}(A_n(\widetilde{f}^*)^c) \leq \tfrac{4SK}{a_n^*}e^{-mb(a_n^*/(2S))} + 2K^{m+1}(2n)^{(d+1)m^2} \frac{e^{-\frac{na_n^{*2}}{128S^2}}}{1 - e^{-\frac{na_n^{*2}}{32S^2}}}.$$

Proof. We use Proposition 10 with $\Phi_{n,k}(f) := vol_n(f_k)$ and $\Phi_k(f) := vol(f_k)$. To apply this proposition we first have to check that Assumption (i) of Proposition 9 is true. To this end we bound

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

Now we use the McDiarmid inequality and the fact that changing one variable X_i only changes $vol_n(g)$ by at most $2S/n$. This implies that for all $g \in \widehat{\mathcal{F}}_n$ and $\epsilon > 0$

$$\mathbb{P}(vol_n(g_k) - vol(g_k) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2S^2}}.$$

The same statement holds for $vol(g_k) - vol_n(g_k)$. This implies Assumption (i) of Proposition 9. Plugging this into Proposition 10 leads to the lemma. □

Lemma 14 (Assumption (3) for Ncut) *Let $f \in \mathcal{F}$ and $g \in \mathcal{F}_n$. Then we have*

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

Proof. We begin with the following inequalities :

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

and

$$\begin{aligned}
|vol(f_k) - vol(g_k)| &= |\int\int f_k(X)s(X,Y) - g_k(X)s(X,Y)| \\
&\leq S \int\int_{\{f=g\}} 0 + S \int\int_{(\{f=g\})^c} 1 \\
&= Sd(f,g) \\
&\leq 2Sd(f,g).
\end{aligned}$$

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

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

On the other hand if $vol(g_k) = 0$ then we have $|cut(f_k)| \leq |vol(f_k)| \leq 2Sd(f,g)$ so the following hold true :

$$\frac{cut(f_k)}{vol(f_k)} - \frac{cut(g_k)}{vol(\widetilde{f}_k)} = \frac{cut(f_k)}{vol(f_k)} \leq \frac{2Sd(f,g)}{a} \leq \frac{4S}{a} d(f,g).$$

So all in all we have

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

We can use the same technique to bound $Ncut(g) - Ncut(f)$. This proves the lemma. □

Now we can state a consistency theorem and a rate of convergence.

Theorem 15 (Consistency of nearest neighbor clustering with Ncut) *Assume that the similarity measure s is bounded by S, $a_n > a, a_n \to a, m \to \infty$ and*

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

Then $Ncut$ is universally weakly consistent, that is for all probability measure $Ncut(f_n)$ tends to $Ncut(f^*)$ in probability.

Proof. The general assumptions of Theorem 1 and Assumption (4) are clearly satisfied. Lemma 14 gives Assumption (3). Now for the Assumption (1) and (2) we have to check some limits.

First of all remark that the assumptions $a_n \to a, m \to \infty$ and $\frac{m^2 \log n}{n(a-a_n)^2} \to 0$ implies that $n(a-a_n)^2 \to \infty$ and $\frac{m^2 \log n}{n} \to 0$ and $\frac{m^2 \log n}{n a_n^{*2}} \to 0$ (recall Lemma 8).

Because of Lemma 12 to check Assumption (1) we only have to prove that

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

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

Because of Lemma 13 and since $m \to \infty$, to check Assumption (2) we only have to prove that

$$K^m(2n)^{(d+1)m^2} e^{-n(a_n-a)^2} \to 0 \text{ and } K^m(2n)^{(d+1)m^2} e^{-n a_n^{*2}} \to 0.$$

It is easy to see that $K^m(2n)^{(d+1)m^2} e^{-n(a_n-a)^2} = e^{-n(a-a_n)^2\left(\frac{(m+1)\log K + (d+1)m^2 \log(2n)}{-n(a-a_n)^2} + 1\right)}$ so using that $\frac{m^2 \log n}{n(a-a_n)^2} \to 0$ and $n(a-a_n)^2 \to \infty$ we have proved the first limit. For the second one we use $\frac{m^2 \log n}{n a_n^{*2}} \to 0$ and $n a_n^{*2} \to \infty$ since $a_n^* \to a^* > 0$ (recall Lemma 8). □

Theorem 16 (Rate of Convergence of nearest neighbor clustering with Ncut) *Let N and a_n^* be defined as in Lemma 8 . Then for $n \geq N$ the following bound holds true:*

$$\begin{aligned}
&\mathbb{P}(|Ncut(f_n) - Ncut(f^*)| \geq \epsilon) \\
&\leq 2K^{m+1}(2n)^{(d+1)m^2} \left(\frac{4e^{-\frac{na^2\epsilon^2}{2048S^2K^2}}}{1-4Ke^{-\frac{na^2\epsilon^2}{512S^2K^2}}} + \frac{e^{-\frac{n(a_n-a)^2}{32S^2}}}{1-e^{-\frac{n(a_n-a)^2}{8S^2}}} + \frac{e^{-\frac{na_n^{*2}}{128S^2}}}{1-e^{-\frac{na_n^{*2}}{32S^2}}} \right) \\
&+ \frac{4SK}{a_n^*} e^{-mb(a_n^*/(2S))} + \frac{16SK}{a\epsilon} e^{-mb(a\epsilon/(8SK))}.
\end{aligned}$$

Proof. Since the general assumptions of Theorem 1 hold and Assumption (3) of Theorem 1 is true according to Lemma 14 we can apply Theorem 7. The rate of convergence then follows from Lemmas 12, 13 and 14. □

Remark *The conditions on a_n and m are easily satisfied as soon as they do not converge too fast to their limit. For example, if we define*

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

then

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

Moreover, in this case we can also see that the rate of convergence given by Theorem 16 is valid as soon as $n > e^{1/a^*}$ (where a^* is given as in Lemma 8).

4.2 Ratio cut objective function

In this section we will prove the consistency of nearest neighbor clustering for the Ratiocut objective function, which is the objective function under consideration by unnormalized spectral clustering. For the sake of shortness we refrain from proving rates of for this case, which can be done analogous to the Ncut case. Again we will assume that the similarity function s is upper bounded by a constant $S > 0$.

The Ratiocut objective function will be denoted by RC. For a clustering function f we define

$$n_k := \frac{1}{n}\sum_{i=1}^{n} f_k(X_i)$$

$$RC_n(f) := \sum_{k=1}^{K} \frac{cut_n(f_k)}{n_k(f)}$$

$$RC(f) := \sum_{k=1}^{K} \frac{cut(f_k)}{\mathbb{E}f_k(X)}$$

and consider the properties

$$A_n(f) = true \iff \forall k \in \{1,\dots,K\}, n_k(f) > a_n$$
$$A(f) = true \iff \forall k \in \{1,\dots,K\}, \mathbb{E}f_k(X) > a.$$

Lemma 17 (Assumption (1) for Ratiocut) *For $f \in \widetilde{\mathcal{F}}_n$ we have*

$$\mathbb{P}(|RC_n(f) - RC(f)| > \epsilon) \leq 2Ke^{-\frac{na^2\epsilon}{8S^2K^2}} + 2Ke^{-\frac{na^2\epsilon^2}{2K^2}}.$$

Proof. Using exactly the same proof as for Lemma 12 (just changing $vol_n(f_k)$ to n_k and $vol(f_k)$ to $\mathbb{E}f_k(X)$ and using the fact that $cut(f_k) \leq S\mathbb{E}f_k(X)$) we get

$$\mathbb{P}(|RC_n(f) - RC(f)| > \epsilon)$$

$$\leq K\sup_k \left(\mathbb{P}(|cut_n(f_k) - cut(f_k)| > \frac{a}{(S+1)K}\epsilon) + \mathbb{P}(|n_k(f) - \mathbb{E}f_k(X)| > \frac{a}{(S+1)K}\epsilon) \right).$$

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

Lemma 18 (Assumption (2) for Ratiocut) *If $a_n > a, a_n \to a$ and*

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

then Assumption (2) of Theorem 1 is satisfied.

Proof. We will use Proposition 9 with $\Phi_{k,n}(f) = n_k(f)$ and $\Phi_k(f) = \mathbb{E}f_k(X)$. Clearly Assumption (i) of Proposition 9 is satisfied since $|\Phi_k(f) - \Phi_k(g)| = |\mathbb{E}f_k(X) - \mathbb{E}g_k(X)| \leq d(f,g)$.

For Assumptions (ii) and (iii) of Proposition 9 we apply the McDiarmid inequality to get that for all $g \in \widehat{\mathcal{F}}_n$ and $\epsilon > 0$:

$$\mathbb{P}(n_k(g) - \mathbb{E}g_k(X)) \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

Then we apply the same methods as in the proof of Theorem 15. □

Lemma 19 (Assumption (3) for Ratiocut) *Let $f \in \mathcal{F}$ and $g \in \mathcal{F}_n$. Then we have:*

$$|RC(f) - RC(g)| \leq \frac{2(S+1)K}{a} d(f,g).$$

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

Theorem 20 (Consistency of nearest neighbor clustering with Ratiocut) *Assume that the similarity measure s is bounded by S, $a_n > a, a_n \to a, m \to \infty$ and*

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

Then Ratiocut is universally weakly consistent, that is for all probability measure $RC(f_n)$ tends to $RC(f^*)$ in probability.

4.3 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 their ratio, and this is what we want to consider in this section. The within- and between-cluster similarities of cluster k are given as

$$\begin{aligned}
\mathrm{I}_n(f_k) &:= \frac{1}{n(n-1)}\sum_{i\neq j} f_k(X_i)f_k(X_j)s(X_i,X_j) = vol_n(f_k) - cut_n(f_k)\\
\mathrm{B}_n(f_k) &:= \frac{1}{n(n-1)}\sum_{i\neq j} f_k(X_i)(1-f_k(X_j))s(X_i,X_j) = cut_n(f_k)\\
\mathrm{I}(f_k) &:= vol(f_k) - cut(f_k)\\
\mathrm{B}(f_k) &:= cut(f_k)
\end{aligned}$$

The ratio of between- and within cluster similarity BWR is now defined as

$$\mathrm{BWR}_n(f) = \sum_{k=1}^{K}\frac{\mathrm{B}_n(f_k)}{\mathrm{I}_n(f_k)} \qquad\qquad \mathrm{BWR}(f) = \sum_{k=1}^{K}\frac{\mathrm{B}(f_k)}{\mathrm{I}(f_k)}.$$

In this section we consider the properties

$$\begin{aligned}
A_n(f) = true &:\Longleftrightarrow \forall k\in\{1,\dots,K\}, \mathrm{I}_n(f_k) > a_n\\
A(f) = true &:\Longleftrightarrow \forall k\in\{1,\dots,K\}, \mathrm{I}(f_k) > a.
\end{aligned}$$

Lemma 21 (Assumption (1) for BWR **)** *For $f\in\widetilde{\mathcal{F}}_n$ we have*

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

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

$$\begin{aligned}
\tfrac{\mathrm{B}_n(f_k)}{\mathrm{I}_n(f_k)} - \tfrac{\mathrm{B}(f_k)}{\mathrm{I}(f_k)} &= \tfrac{\mathrm{I}(f_k)\,\mathrm{B}_n(f_k) - \mathrm{I}_n(f_k)\,\mathrm{B}(f_k)}{\mathrm{I}_n(f_k)\,\mathrm{I}(f_k)}\\
&\leq \tfrac{\mathrm{I}(f_k)(\mathrm{B}(f_k)+\epsilon) - (\mathrm{I}(f_k)-\epsilon)\,\mathrm{B}(f_k)}{\mathrm{I}_n(f_k)\,\mathrm{I}(f_k)}\\
&= \tfrac{\epsilon}{\mathrm{I}_n(f_k)}\,\tfrac{\mathrm{I}(f_k)+\mathrm{B}(f_k)}{\mathrm{I}(f_k)}\\
&\leq \tfrac{2S\epsilon}{a^2}
\end{aligned}$$

The analogous statement holds for $\frac{\mathrm{B}(f_k)}{\mathrm{I}(f_k)} - \frac{\mathrm{B}_n(f_k)}{\mathrm{I}_n(f_k)}$.
So if $\epsilon \leq S/a$ we have

$$\{|\,\mathrm{I}_n(f_k) - \mathrm{I}(f_k)| \leq a^2\epsilon/(2S)\} \cap \{|\,\mathrm{B}_n(f_k) - \mathrm{B}(f_k)| \leq a^2\epsilon/(2S)\} \subset \{\left|\frac{\mathrm{B}_n(f_k)}{\mathrm{I}_n(f_k)} - \frac{\mathrm{B}(f_k)}{\mathrm{I}(f_k)}\right| \leq \epsilon\}.$$

Now if $\epsilon \leq SK/a$ we have

$$\begin{aligned}
&\mathbb{P}(|\,\mathrm{BWR}_n(f) - \mathrm{BWR}(f)| > \epsilon)\\
&\leq K\sup_k \mathbb{P}\left(\left|\tfrac{\mathrm{B}_n(f_k)}{\mathrm{I}_n(f_k)} - \tfrac{\mathrm{B}(f_k)}{\mathrm{I}(f_k)}\right| > \epsilon/K\right)\\
&\leq K\sup_k \left(\mathbb{P}(|\,\mathrm{I}_n(f_k) - \mathrm{I}(f_k)| > a^2\epsilon/(2SK)) + \mathbb{P}(|\,\mathrm{B}_n(f_k) - \mathrm{B}(f_k)| > a^2\epsilon/(2SK))\right).
\end{aligned}$$

Using the McDiarmid inequality together with the fact that changing one point changes B_n and I_n by at most $S/(2n)$, we get for $\epsilon \leq SK/a$:

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

On the other hand for $\epsilon > SK/a$ we have

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

So all in all we have proved the lemma. □

Lemma 22 (Assumption (2) for BWR**)** *If $a_n > a, a_n \to a$ and*

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

then Assumption (2) *of Theorem 1 is satisfied.*

Proof. We will check the three assumptions of Proposition 9, using $\Phi_k(f) = \mathrm{I}(f_k)$ and $\Phi_{k,n}(f) = \mathrm{I}_n(f_k)$. To see that Assumption (i) of Proposition 9 holds we use the following inequalities:

$$\begin{aligned}
\mathrm{I}(f_k) - \mathrm{I}(g_k) &= \int\int (f_k(X)f_k(Y) - g_k(X)g_k(Y))s(X,Y) \\
&\leq \int_{\{f=g\}^2} 0 + S\int_{(\{f=g\}^2)^c} 1 \\
&= S(1 - \mathbb{P}(f=g)^2) \\
&= S(1 - (1 - d(f,g))^2) \\
&\leq 2Sd(f,g).
\end{aligned}$$

For Assumptions (ii) and (iii) we get using McDiarmid's inequality that

$$\mathbb{P}(\mathrm{I}_n(f_k) - \mathrm{I}(f_k) \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2S^2}}.$$

Then we apply the same method as in proof of Theorem 15. □

Lemma 23 (Assumption (3) for BWR**)** *Assumption* (3) *of Theorem 1 is satisfied by* BWR .

Proof. Let $\epsilon > 0$, $f \in \mathcal{F}$ and $g \in \mathcal{F}_n$. We have already proved the two following inequalities (in the proofs of Lemmas 14 and 22) :

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

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

Now if $2Sd(f,g) \leq a/2$ then using that $\mathrm{I}(f_k) > a$ and we get $\mathrm{I}(g_k) \geq a/2 > 0$. By the same technique as at the beginning of Lemma 21 we get

$$|\operatorname{BWR}(f) - \operatorname{BWR}(g)| \leq \frac{2SK}{a^2} 2Sd(f,g).$$

To rewrite it we have :

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

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

If $\epsilon \leq SK/a$ then we have :

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

On the other hand if $\epsilon > SK/a$ then

$$d(f,g) \leq \frac{a}{4S} \Rightarrow |\operatorname{BWR}(f) - \operatorname{BWR}(g)| \leq \frac{4S^2K}{a^2} d(f,g) \leq SK/a \leq \epsilon$$

so we have proved the lemma. □

Theorem 24 (Consistency of nearest neighbor clustering with BWR**)** *Assume that the similarity measure s is bounded by S, $a_n > a, a_n \to a, m \to \infty$ and*

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

Then BWR *is universally weakly consistent, that is for all probability measure* $\mathrm{BWR}(f_n)$ *tends to* $\mathrm{BWR}(f^*)$ *in probability.*

4.4 The objective function used by the K-means algorithm

In this section we want to study the objective function used in the K-means algorithm. The standard K-means algorithm tries to minimize the within-cluster sum of squared distances, which we will call WSS. This objective function can be expressed as the squared distances of all data points to the center of their clusters. For a given clustering function f we introduce the following quantities:

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

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

Note that n_k is the same quantity as in Section 4.2.

We would like to point out some important facts. First, as in the previous cases, the empirical quality function is not an unbiased estimator of the true one, that is $\mathbb{E}\,\mathrm{WSS}_n \neq \mathrm{WSS}$ and $\mathbb{E} c_{k,n} \neq c_k$. However, at least we have $\mathbb{E} n_k = \mathbb{E} f_k(X)$ and $\mathbb{E} \frac{1}{n} \sum_{i=1}^{n} f_k(X_i) X_i = \mathbb{E} f_k(X) X$.

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 of the sample, then one can see that the clustering constructed by this minimizer always builds a Voronoi partition with K cells of the data space; the same holds in the limit case. In particular, given the cluster centers $c_{k,n}$ we 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 centers. It may very well happen that a point x belongs to cluster C_i, but in fact is closer to the center $c_{j,n}$ of another cluster C_j than to the center $c_{i,n}$ of its own cluster C_i. This means that 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 that the one of the standard K-means case.

It is due to the problems described above that we have to introduce again the assumption that clusters have a certain minimal size. That is, we consider the properties

$$A_n(f) = true :\iff \forall k \in \{1, \ldots, K\}, n_k(f) > a_n$$
$$A(f) = true :\iff \forall k \in \{1, \ldots, K\}, \mathbb{E} f_k(X) > a.$$

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.

Again we want to verify the conditions of Theorem 1.

Lemma 25 (Assumption (1) for WSS**)** *Assume that* $\operatorname{supp}\mathbb{P} \subset B(0,A) \subset \mathbb{R}^d$ *for some constant* $A > 0$*. For* $f \in \widetilde{\mathcal{F}}_n$ *we have*

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

Proof. First notice that

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

Now we will bound the probability for each of the terms. 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

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

Now we have to take care of the first term which is

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

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 for each coordinate. Before we can do this, we want to show that

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

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

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

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

and similarly for $c_{k,n}^j - c_k^j$.

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 $\frac{1}{n}\sum_{i=1}^{n} f_k(X_i)X_i^j \leq \epsilon$. This implies

$$|c_k^j - c_{k,n}^j| = \frac{1}{n_k}\frac{1}{n}\sum_{i=1}^{n} f_k(X_i)X_i^j \leq \frac{\epsilon}{a} \leq \frac{(A+1)\epsilon}{a}$$

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

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

Together they show that for the coordinate-wise differences

$$\mathbb{P}(|c_k^j - c_{k,n}^j| > \epsilon) \leq 2e^{-\frac{2na^2\epsilon^2}{(A+1)^2}} + 2e^{-\frac{2na^2\epsilon^2}{A^2(A+1)^2}} \leq 4e^{-\frac{2na^2\epsilon^2}{\max(1,A^2)(A+1)^2}}.$$

This leads to

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

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

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

To finish the proof we now combine the probabilities for the first and the second term from the beginning of the proof using a union bound. □

Lemma 26 (Assumption (2) for WSS**)** *We have*

$$\mathbb{P}(A(f_n)^c) \leq 2K^{m+1}(2n)^{(d+1)m^2}\frac{e^{-\frac{n(a_n-a)^2}{8}}}{1 - e^{-\frac{n(a_n-a)^2}{2}}}.$$

If additionally we have $a_n^* > 0$ (recall the definition of a_n^* in Lemma 8) then

$$\mathbb{P}(A_n(\widetilde{f}_k^*)^c \leq a_n) \leq \tfrac{4K}{a_n^*}e^{-mb(a_n^*/2)} + 2K^{m+1}(2n)^{(d+1)m^2}\frac{e^{-\frac{na_n^{*2}}{32}}}{1-e^{-\frac{na_n^{*2}}{8}}}.$$

Proof. We will use the Proposition 10 with $\Phi_k(f) = \mathbb{E}f_k(X)$ and $\Phi_{k,n} = n_k$. Clearly Assumption (i) of Proposition 9 is satisfied since $|\mathbb{E}f_k(X) - \mathbb{E}g_k(X)| \leq d(f,g)$. Moreover, to bound the terms that appear in the bound of Proposition 10 we use the McDiarmid inequality to get that for all $\forall g \in \widehat{\mathcal{F}}$ and $\forall \epsilon > 0$:

$$\mathbb{P}(n_k(g) - \mathbb{E}g_k(X) \geq \epsilon) \leq e^{-2n\epsilon^2}.$$

Plugging this into Proposition 10 proves the lemma. □

Lemma 27 (Assumption (3) for WSS**)** *Assume that* $\operatorname{supp}\mathbb{P} \subset B(0,A)$ *for some constant* $A > 0$*. Let* $f \in \mathcal{F}$ *and* $g \in \mathcal{F}_n$*. Then we have*

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

Proof. 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 $\operatorname{supp}\mathbb{P} \subset B(0,A)$:

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

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

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

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 $|X|R$ 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 the lemma. □

Now we can finally state the consistency theorem and a rate of convergence for nearest neighbor clustering using the WSS objective function. The proofs simply combine the lemmas analogously to the previous sections.

Theorem 28 (Consistency of nearest neighbor clustering with 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* $\operatorname{WSS}(f_n)$ *tends to* $\operatorname{WSS}(f^*)$ *in probability.*

Remark *It is straight forward to see 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 most crucial property is that we need to be able to bound*

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

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

Theorem 29 (Convergence Rate of nearest neighbor clustering using WSS**)** *Assume that* $\operatorname{supp}\mathbb{P} \subset B(0, A)$ *for some constant* $A > 0$*. Let* N *and* a_n^* *be defined as in Lemma 8 . Then for* $n \geq N$ *the following bound holds true:*

$$\mathbb{P}(|\,\mathrm{WSS}(f_n) - \mathrm{WSS}(f^*)| \geq \epsilon)$$

$$\leq 2K^{m+1}(2n)^{(d+1)m^2} \left(\frac{4dKe^{-\frac{na^2\epsilon}{616d\max(1,A^2)A^2(A+1)^2}} + 2e^{-\frac{n\epsilon^2}{32A^4}}}{1-4dKe^{-\frac{na^2\epsilon}{308d\max(1,A^2)A^2(A+1)^2}} - 2e^{-\frac{n\epsilon^2}{8A^4}}} + \frac{Ke^{-\frac{n(a_n-a)^2}{8}}}{1-e^{-\frac{n(a_n-a)^2}{2}}} + \frac{Ke^{-\frac{na_n^{*2}}{32}}}{1-e^{-\frac{na_n^{*2}}{8}}} \right)$$

$$+\tfrac{4K}{a_n^*}e^{-mb(a_n^*/2)} + (16A^2(1+3/a)/\epsilon)e^{-mb(\epsilon/(8A^2(1+3/a)))}.$$

References

S. Bubeck and U. von Luxburg. Overfitting of clustering and how to avoid it. Preprint, 2007.

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

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

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

U. von Luxburg, S. Bubeck, S. Jegelka, and M. Kaufmann. Consistent minimization of clustering objective functions. In *NIPS 2007*, to appear.