# Kernel Normalized Cut: a Theoretical Revisit

**Yoshikazu Terada** [1 2] **Michio Yamamoto** [3 2]

## Abstract

In this paper, we study the theoretical properties of clustering based on the kernel normalized cut. Our first contribution is to derive a nonasymptotic upper bound on the expected distortion rate of the kernel normalized cut. From this result, we show that the solution of the kernel normalized cut converges to that of the population-level weighted $k$-means clustering on a certain reproducing kernel Hilbert space (RKHS). Our second contribution is the discover of the interesting fact that the population-level weighted $k$-means clustering in the RKHS is equivalent to the population-level normalized cut. Combining these results, we can see that the kernel normalized cut converges to the population-level normalized cut. The criterion of the population-level normalized cut can be considered as an indivisibility of the population distribution, and this criterion plays an important role in the theoretical analysis of spectral clustering in Schiebinger et al. (2015). We believe that our results will provide deep insights into the behavior of both normalized cut and spectral clustering.

## 1. Introduction

Data clustering is one of the most important problems in unsupervised learning and has many applications. Given a set of data points, the purpose of clustering algorithms is to discover groups (clusters) of data points based on some sense of similarity. In the context of the clustering problem, we do not have any information or training labels about these groups and, therefore, the sense (or type of optimality) of the obtained groups is very important. Usually, we assume that the data points $X_1, \ldots, X_n$ are independently drawn from an unknown underlying distribution, called the population distribution. In this setting, as mentioned in von Luxburg et al. (2008), it is important to check whether a clustering algorithm satisfies the following requirements:

- When the sample size $n$ goes to infinity, the clusterings constructed by the clustering algorithm converge to a certain clustering of the underlying data space corresponding to the population distribution.
- If so, the limit clustering of the underlying space provides a partition of the underlying space that is optimal in some sense for the population distribution.

For example, $k$-means clustering satisfies these requirements under very mild conditions (Pollard, 1981) and can reach a fast rate of convergence under the margin condition (Levrard, 2015). Even though classical clustering algorithms such as $k$-means clustering have such nice properties, the obtained partitions are too simple and sometime do not capture group structures that are reasonably intuitive. In recent years, graph partitioning approaches represented by spectral clustering have drawn much attention (Shi & Malik, 2000; Ng et al., 2002; Rosasco et al., 2010; Cao & Chen, 2011; Arias-Castro et al., 2012; Schiebinger et al., 2015; Gracía Trillos et al., 2016; Davis & Sethuraman, 2016; Gracía Trillos & Slepčev, 2016). When we use graph partitioning approaches to cluster multivariate data, we treat the similarity matrix between the data points as an adjacency matrix of a graph and then apply a graph cut algorithm to the similarity matrix. Typically, the similarity measure depends on the distances between data points and the connectivity length scale parameter $\epsilon_n$, which determines the length scale of the edges.

Spectral clustering was introduced in the machine learning community with applications to image segmentation and clustering multivariate data (e.g., Shi & Malik, 2000; Ng et al., 2002), and it has been successfully applied in various fields. The input of graph-based clustering algorithms such as spectral clustering is a similarity matrix between the data points. Since spectral clustering is one of modern state-of-the-art clustering algorithms, there have been several studies addressing its theoretical properties. For example, von Luxburg et al. (2008), Rosasco et al. (2010) and Cao & Chen (2011) established the consistency of

[1]Graduate School of Engineering Science, Osaka University, Osaka, Japan [2]RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan [3]Graduate School of Environmental and Life Science, Okayama University, Okayama, Japan. Correspondence to: Yoshikazu Terada <terada@sigmath.es.osaka-u.ac.jp>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. 

spectral clustering with a fixed connectivity length scale. Recently, Schiebinger et al. (2015) discovered additional detailed properties concerning the performance of spectral clustering, from which we can see why spectral clustering works so well. For the case where the connectivity length scale $\epsilon_n$ goes to zero as the sample size goes to infinity, Gracía Trillos & Slepčev (2016) established the consistency of spectral clustering.

The roots of spectral clustering are spectral graph partitioning (Donath & Hoffman, 1973; Fideler, 1973) and spectral clustering is a relaxation of the graph cut called the normalized cut (Ncut for short; Shi & Malik, 2000). However, in contrast to spectral clustering, Ncut has not received much attention since the search for the best graph cut in the sense of Ncut is an NP-hard problem. Fortunately, Dhillon et al. (2007) found an efficient algorithm to obtain the local optimal solution of the normalized cut. Several theoretical results of Ncut have been established. Under the case where the connectivity length scale $\epsilon_n$ goes to zero, Arias-Castro et al. (2012) show the consistency of restricted normalized cut that is minimized over a particular family of subsets of the data points. Moreover, Gracía Trillos et al. (2016) improved these results of Ncut. Specifically, they established the consistency of a non-restricted normalized cut that is minimized over all possible partitions of the data points.

Even though using a small $\epsilon_n$ has an advantage from the viewpoint of computational cost and could provide better resolution, we also often use a fixed connectivity length scale $\epsilon > 0$ in practical situations. In this paper, we focus on a setting with a fixed connectivity length scale $\epsilon > 0$ as in von Luxburg et al. (2008) and Schiebinger et al. (2015). Ben-David et al. (2006) briefly described the stability of Ncut in this setting. As our first contribution, we provide a more detailed consistency result showing that the optimal cut of the empirical normalized cut converges to the optimal solution of the population-level weighted $k$-means clustering on a certain reproducing kernel Hilbert space (RKHS) when we use a kernel function as a similarity measure. Note that, in Dhillon et al. (2007) and Ben-David et al. (2006), they did not deal with the corresponding RKHS. Moreover, we derive a nonasymptotic upper bound on the expected distortion rate of the kernel Ncut. We also show that, under a specific condition for the population distribution, the kernel Ncut with the true degree function can reach a fast rate convergence. This result corresponds to an extension of the $k$-means result in Levrard (2015) for the weighted $k$-means. On the RKHS, we can obtain the clear optimality of the partition of the clustering. However, the optimality of the partition in the original data space is still not clear. Our second contribution is to discover an interesting fact that the weighted $k$-means clustering in the RKHS is equivalent to the population-level Ncut over partitions of the data space. Combining these results, we can conclude that the partition by the empirical Ncut with a kernel function converges to the optimal partition of the population-level Ncut as the sample size goes to infinity. An overview of our study is shown in Figure 1.

This paper is organized as follows. In Section 2, the notation and some properties of Ncut are introduced. Section 3 establishes the consistency of the kernel Ncut and provides a nonasymptotic upper bound on the expected distortion rate of the kernel Ncut. In Section 4, it is shown that the weighted $k$-means clustering in such RKHS is equivalent to the population-level Ncut over partitions of the data space. In Section 5, we will look at the difference between spectral clustering and Ncut through the numerical experiments. The last section concludes the paper with some discussion.

## 2. Preliminaries

Let $\mathcal{X}$ be a compact metric data space with the Borel $\sigma$-algebra $\mathscr{B}$, and let $\mathbb{P}$ be a probability measure on $(\mathcal{X}, \mathscr{B})$. Throughout this paper, we assume that $\mathcal{X}$ coincides with the support of $\mathbb{P}$ without loss of generality. We will denote by $X_1, \ldots, X_n$ sample points independently drawn from $\mathbb{P}$ and by $\mathbb{P}_n$ the empirical measure based on $X_1, \ldots, X_n$. Write $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ denote a symmetric, continuous, positive definite kernel function. We will denote by $\mathcal{H}_k$ the RKHS with reproducing kernel $k$. Let $\langle \cdot, \star \rangle_{\mathcal{H}_k}$ and $\| \cdot \|_{\mathcal{H}_k}$ denote the inner product and the norm of $\mathcal{H}_k$, respectively. For kernel function $k$, the canonical feature map is defined by $\psi_k : \mathcal{X} \to \mathcal{H}_k$ by $\psi_k(x) := k(x, \cdot)$ $(x \in \mathcal{X})$.

We suppose that $k$ is bounded away from zero and infinity, that is, $\exists c_L, c_U > 0$; $\forall x, y \in \mathcal{X}$; $c_L < k(x, y) < c_U$. For sample points $X_1, \ldots, X_n$ and a fixed kernel function $k$ as in the above general assumption, we will write the kernel matrix as $K_n = (k_{ij})_{n \times n} := (k(X_i, X_j))_{n \times n}$. Let us denote by $\hat{d}_n(\cdot) := \int_{\mathcal{X}} k(\cdot, x) \mathbb{P}_n(dx) = \sum_{j=1}^n k(\cdot, X_j)/n$ the empirical degree function, and write $D_n := \mathrm{diag}(\hat{d}_n(X_1), \ldots, \hat{d}_n(X_n))$. When we consider the kernel matrix $K_n$ as an adjacency matrix of a weighted graph for which the vertices are data points, we can use Ncut for clustering data points.

For a given $M$-partition $\mathscr{P}_M := \{W_1, \ldots, W_M\}$ of data set $\mathcal{X}_n := \{X_1, \ldots, X_n\}$, the objective function of Ncut is defined as

$$\begin{aligned} \mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}_n) &:= \sum_{m=1}^M \frac{\mathrm{Mcut}(W_m, \mathcal{X}_n \setminus W_m)}{\mathrm{vol}(W_m)} \\ &= M - \sum_{m=1}^M \frac{\mathrm{Mcut}(W_m, W_m)}{\mathrm{vol}(W_m)} \end{aligned}$$

where $\mathrm{Mcut}(A, B) := \sum_{i \in A} \sum_{j \in B} k_{ij}$ is the objective function of the minimum cut and $\mathrm{vol}(A) := \sum_{i \in A} \hat{d}_n(X_i)$.

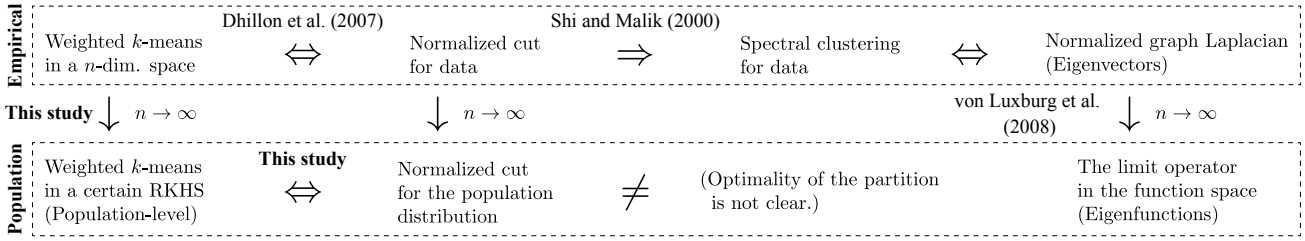


*Figure 1.* Overview of our study with its related results.

Using the membership matrix $U_n = (u_{im})_{n\times M} := (\mathbb{1}(i \in W_m))_{n\times M}$, we can rewrite $\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}_n)$ as the following matrix formula:

$$\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}_n) = M - \mathrm{tr}\{\tilde{U}_n^T D_n^{-1/2} K_n D_n^{-1/2} \tilde{U}_n\}, \tag{1}$$

where $\tilde{U}_n := D_n^{1/2} U_n (U_n^T D_n U_n)^{-1/2}$. Since $\tilde{U}_n^T \tilde{U}_n = I_M$, we have the following relationship

$$\begin{aligned}\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}_n) &= \mathrm{tr}\{\tilde{U}_n^T (I_n - D_n^{-1/2} K_n D_n^{-1/2}) \tilde{U}_n\} \\ &= \mathrm{tr}\{\tilde{U}_n^T L_n \tilde{U}_n\},\end{aligned}$$

where $L_n := D_n^{-1/2}(D_n - K_n)D_n^{-1/2}$ is the normalized graph Laplacian. When $\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}_n)$ is considered as the function of the membership matrix $U_n$, we will denote it by $\mathrm{Ncut}(U_n \mid \mathbb{P}_n)$. Since the minimization problem of $\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}_n)$ with respect to $\mathscr{P}_M$ (or equivalently $U_n$) is NP-hard, we relax this problem to the following problem in spectral clustering:

$$\min_{\tilde{U}^T \tilde{U} = I_M} \mathrm{tr}\{\tilde{U}^T L_n \tilde{U}\}.$$

For details about spectral clustering, we refer the reader to von Luxburg (2007).

Now, we briefly explain the relationship, discovered by Dhillon et al. (2007), between Ncut and the weighted kernel $k$-means (WKKM). The objective function of the weighted kernel $k$-means using the kernel function $h$ is given by

$$\begin{aligned}&\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}_n) \\ &:= \frac{1}{n}\sum_{i=1}^n w_i \min_{1\le m\le M} \|\psi_h(X_i) - \mu_m\|_{\mathcal{H}_h}^2,\end{aligned}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$ is a codebook in $\mathcal{H}_h^M$, and $w_i$ is the weight of the $i$-th data point. Let $U_n := (u_{im})_{n\times M}$ be a binary membership matrix that specifies the cluster membership for each data point. Without loss of generality, we only consider a binary membership matrix $U_n$ such that $\forall m \in \{1, \ldots, M\};\ \exists i \in \{1, \ldots, n\},\ u_{im} = 1$. If $u_{im} = 0$ for all $i \in \{1, \ldots, n\}$, we remove the $m$-th column from $U_n$. For a given $U_n$, the optimal mean $\mu_m$ can be given by

$$\hat{\mu}_m := \frac{\sum_{i=1}^n w_i u_{im} \psi_h(X_i)}{\sum_{i=1}^n w_i u_{im}}.$$

Using this fact and the reproducing property, we can rewrite $\mathrm{WKKM}(\boldsymbol{\mu} \mid \mathbb{P}_n)$ as a function of $U_n$ such that

$$\begin{aligned}&\mathrm{WKKM}_h(U_n \mid \mathbb{P}_n) \\ &:= \frac{1}{n}\sum_{i=1}^n w_i h_{ii} - \frac{1}{n}\sum_{m=1}^M \sum_{i=1}^n \sum_{j=1}^n u_{im} u_{jm} w_i w_j h_{ij} / s_m \\ &= \frac{1}{n}\left[\mathrm{tr}\left(W_n^{1/2} H_n W_n^{1/2}\right) - \mathrm{tr}\left(\tilde{U}_n^T W_n^{1/2} H_n W_n^{1/2} \tilde{U}_n\right)\right],\end{aligned} \tag{2}$$

where $H_n := (h_{ij})_{n\times n} := (h(X_i, X_j))_{n\times n}$, $W_n := \mathrm{diag}(w_1, \ldots, w_n)$, $S := \mathrm{diag}(s_1, \ldots, s_M) = U_n^T W_n U_n$, and $\tilde{U}_n := W_n^{1/2} U_n S^{-1/2}$. The first term in the box brackets is not related to $U_n$ and therefore can be ignored during optimization. Therefore, from (2), we can see that Ncut defined by (1) coincides with WKKM with $W_n := D_n$ and $H_n := D_n^{-1} K_n D_n^{-1}$. Let $\tilde{d}_n(x) := \sum_{i=1}^n k(\cdot, x)$ and $\tilde{D}_n := \mathrm{diag}(\tilde{d}_n(X_1), \ldots, \tilde{d}_n(X_n))$. It is worth noting that, in Dhillon et al. (2007), they discovered the above equivalence with $H_n := \tilde{D}_n^{-1} K_n \tilde{D}_n^{-1}$ and $W_n := \tilde{D}_n$. In the sense of the optimization problem, there is no difference between these two representations. However, to consider the asymptotic properties of the kernel Ncut, our modified representation with $W_n := D_n$ and $H_n := D_n^{-1} K_n D_n^{-1}$ is useful. Let $\hat{h}_n : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be defined as

$$\hat{h}_n(x, y) = \frac{k(x, y)}{\hat{d}_n(x)\hat{d}_n(y)} \quad (x, y \in \mathcal{X}).$$

Then, it follows that the empirical kernel Ncut is equivalent to the weighted $k$-means on the RKHS $\mathcal{H}_{\hat{h}_n}$ with kernel $\hat{h}_n$. However, the kernel function $\hat{h}_n$ depends on the data $\mathcal{X}_n$, and the corresponding RKHS $\mathcal{H}_{\hat{h}_n}$ changes with sample size $n$ and data $\mathcal{X}_n$. Thus, it is difficult to consider the asymptotic properties of $\mathrm{WKKM}_{\hat{h}_n}(U_n)$ directly.

## 3. Consistency of the kernel normalized cut

The equivalence described in the previous section between the kernel normalized cut and the weighted kernel $k$-means is from the view point of optimization problem. Now, through a theoretical analysis of the weighted kernel $k$-means, we consider the large sample limit of the empirical normalized cut.

### 3.1. Basic consistency results

We define the degree function $d : \mathcal{X} \to \mathbb{R}$ by $d(\cdot) := \int_{\mathcal{X}} k(\cdot, x)\, \mathbb{P}(dx)$. To develop the theoretical properties of Ncut, we introduce a (non-random) kernel function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$h(x, y) := \frac{k(x, y)}{d(x)d(y)} \quad (x, y \in \mathcal{X}).$$

The population level objective function of WKKM with kernel $h$ is defined by

$$\begin{aligned}&\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P}) \\ &:= \int d(x) \min_{1 \le m \le M} \|\psi_h(x) - \mu_m\|^2_{\mathcal{H}_h}\, \mathbb{P}(dx).\end{aligned}$$

The following lemma provides the fundamental connection between the Ncut algorithm proposed by Dhillon et al. (2007) and the weighted kernel $k$-means $\mathrm{WKKM}_h(U_n)$ with $w_i = d(X_i)$ on RKHS $\mathcal{H}_h$, which does not depend on $n$. In addition, we provide a non-asymptotic bound for the uniform difference between the empirical and true degree functions.

**Lemma 1.** *Assume the general assumption described in Section 2. Then, we have*

$$\max_{U_n} |\mathrm{Ncut}(U_n \mid \mathbb{P}_n) - \mathrm{WKKM}_h(U_n)| \le \frac{2c_U}{c_L^2} \|\hat{d}_n - d\|_\infty,$$

*where* $\|\hat{d}_n - d\|_\infty := \sup_{x \in \mathcal{X}} \left|\hat{d}_n(x) - d(x)\right|$. *Moreover, for all* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*, we have*

$$\|\hat{d}_n - d\|_\infty \le 2\frac{c_U^{3/2}}{\sqrt{n}} + (c_U - c_L)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

*Proof.* See Section A.1 in the supplementary material. □

Let $\hat{U}_n = (\hat{u}_{im})_{n \times M}$ be the estimated membership matrix by the empirical Ncut $\mathrm{Ncut}(\cdot \mid \mathbb{P}_n)$ with the kernel matrix $K_n$, and define the corresponding codebook $\tilde{\boldsymbol{\mu}}_n = (\tilde{\mu}_1^{(n)}, \ldots, \tilde{\mu}_M^{(n)})$ by

$$\tilde{\mu}_m^{(n)} = \frac{\sum_{i=1}^n \hat{u}_{im} \hat{d}_n(X_i) \psi_{\hat{h}_n}(X_i)}{\sum_{i=1}^n \hat{d}_n(X_i) \hat{u}_{im}} \in \mathcal{H}_{\hat{h}_n}.$$

We expect that $\tilde{\boldsymbol{\mu}}_n$ will perform close to optimal for the population distribution $\mathbb{P}$, that is, $\mathrm{WKKM}_h(\tilde{\boldsymbol{\mu}}_n \mid \mathbb{P}) \approx \min_{\boldsymbol{\mu} \in \mathcal{H}_h} \mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{P})$. However, since $\tilde{\mu}_m^{(n)}$ is not an element of $\mathcal{H}_h$ in general, we cannot evaluate the distortion $\mathrm{WKKM}_h(\cdot \mid \mathbb{P})$ at $\tilde{\boldsymbol{\mu}}_n$. To overcome this difficulty, we define the following codebook $\hat{\boldsymbol{\mu}}_n = (\hat{\mu}_1^{(n)}, \ldots, \hat{\mu}_M^{(n)})$ associated with $\hat{U}_n$ as

$$\hat{\mu}_m^{(n)} = \frac{\sum_{i=1}^n \hat{u}_{im} d(X_i) \psi_h(X_i)}{\sum_{i=1}^n d(X_i) \hat{u}_{im}} \in \mathcal{H}_h.$$

The following lemma provides a uniform upper bound of the difference between $\tilde{\boldsymbol{\mu}}_n$ and $\hat{\boldsymbol{\mu}}_n$.

**Lemma 2.** *For any* $\hat{U}_n$ *and all* $m \in \{1, \ldots, M\}$*,*

$$\|\tilde{\mu}_m^{(n)} - \hat{\mu}_m^{(n)}\|_\infty \le \frac{c_U^2}{c_L^2} \|\hat{d}_n - d\|_\infty + \frac{c_U}{c_L} \|\hat{d}_n - d\|_\infty.$$

The proof of this lemma is nearly the same as that of Lemma 1.

From Lemma 1 and Borel-Cantelli lemma, we have $\|\hat{d}_n - d\|_\infty \to 0$ $a.s.$ as $n \to \infty$. Therefore, the difference between $\tilde{\boldsymbol{\mu}}_n$ and $\hat{\boldsymbol{\mu}}_n$ converges to zero almost surely. Accordingly, we primarily focus on the behavior of the codebook $\hat{\boldsymbol{\mu}}_n$ instead of $\tilde{\boldsymbol{\mu}}_n$.

For simplicity of notation, we introduce the following function:

$$\zeta = \begin{cases} \mathcal{H}_h^M \times \mathcal{X} \to \mathbb{R}, \\ (\boldsymbol{\mu}, x) \mapsto d(x) \min_{1 \le m \le M} \|\psi_h(x) - \mu_m\|^2_{\mathcal{H}_h}. \end{cases}$$

Here, the norm $\|\cdot\|$ of any element $\boldsymbol{\mu} \in \mathcal{H}_h^M$ is defined as $\|\boldsymbol{\mu}\| = \sqrt{\sum_{m=1}^M \|\mu_m\|^2_{\mathcal{H}_h}}$. For a distribution $\mathbb{Q}$, $\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{Q})$ can take the form $\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}[\zeta(\boldsymbol{\mu}, X)]$, where $X \sim \mathbb{Q}$ and $\mathbb{E}_{\mathbb{Q}}[\cdot]$ is the expectation with respect to $\mathbb{Q}$. It is assumed that the support of $\mathbb{P}$ has more than $M$ points. For $c \in \mathcal{H}_h$ and $r > 0$, we will denote by $\mathcal{B}(c, r)$ the closed ball with center $c$ and radius $r$, that is, $\mathcal{B}(c, r) := \{f \in \mathcal{H}_h \mid \|f - c\|_{\mathcal{H}_h} \le r\}$. Let $D := \sqrt{c_U}/c_L$. By the assumption on the kernel function $k$, we have $\|\psi_h(x)\|_{\mathcal{H}_h} \le D$ for all $x \in \mathcal{X}$. From Proposition 5 described later, there are minimizers $\boldsymbol{\mu}^*$ of $\mathrm{WKKM}_h(\cdot \mid \mathbb{P})$ on $\mathcal{B}(0, D)^M \subset \mathcal{H}_h^M$. We will denote by $\mathcal{M}$ the set of minimizers of the risk $\mathrm{WKKM}_h(\cdot \mid \mathbb{P})$. The loss $\ell(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*)$ is defined as

$$\begin{aligned}\ell(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) &:= \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n \mid \mathbb{P}) - \mathrm{WKKM}_h(\boldsymbol{\mu}^* \mid \mathbb{P}) \\ &\ge 0.\end{aligned}$$

The following proposition provides a nonasymptotic bound for the loss $\ell(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*)$.

**Proposition 3.** *Assume the general assumption described in Section 2. Suppose that the number of elements in the*

*support of $\mathbb{P}$ is greater than the given number of clusters, that is, $\#(\mathrm{supp}(\mathbb{P})) \geq M$. Then, for all $\delta \in (0,1)$, with probability at least $1-2\delta$, we have*

$$\begin{aligned}\ell(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) &= \mathrm{WKKM}_h(\hat{\boldsymbol{\mu}}_n \mid \mathbb{P}) - \mathrm{WKKM}_h(\boldsymbol{\mu}^* \mid \mathbb{P}) \\ &\leq \frac{4\sqrt{c_U}D\{\sqrt{c_U}MD + 2(M + Dc_U)\}}{\sqrt{n}} \\ &\quad + 4D^2(2c_U - c_L)\sqrt{\frac{\log(2/\delta)}{2n}}.\end{aligned}$$

*Proof.* See Section A.2 in the supplementary material. □

Therefore, along with the fact described in Section 2, we can see that the empirical Ncut with the kernel function $k$ converges to the population-level weighted $k$-means clustering on the RKHS $\mathcal{H}_h$.

**Theorem 4.** *Assume the general assumption described in Section 2. Suppose that the number of elements in the support of $\mathbb{P}$ is greater than the given number of clusters, that is, $\#(\mathrm{supp}(\mathbb{P})) \geq M$. Then, we have*

$$\lim_{n\to\infty} \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*(\hat{\boldsymbol{\mu}}_n)\| = 0 \ \ a.s.,$$

*where $\boldsymbol{\mu}^*(\boldsymbol{\mu}) \in \arg\min_{\boldsymbol{\mu}^* \in \mathcal{M}} \|\boldsymbol{\mu} - \boldsymbol{\mu}^*\|$.*

**Proof sketch.** From Proposition 3, we have that, for any $\epsilon > 0$,

$$\mathbb{P}(\ell(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) > \epsilon) \leq \exp(-n\epsilon^2/Const.),$$

where $Const.$ is a constant depending on $c_L$, $c_U$, and $M$. According to the Borel-Cantelli lemma, we can obtain almost sure convergence $\ell(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) \to 0$ $a.s.$ as $n \to \infty$. Combining this and the continuity of the risk $\mathrm{WKKM}_h(\cdot \mid \mathbb{P})$, we can prove the claim in much the same way as the proof of Pollard (1981). □

Combining this with Lemma 2, we also have

$$\|\tilde{\boldsymbol{\mu}}_n - \boldsymbol{\mu}^*(\hat{\boldsymbol{\mu}}_n)\|_\infty := \max_{1\leq m\leq M} \|\tilde{\mu}_m^{(n)} - \mu_m^*\|_\infty \to 0$$

almost surely as $n \to \infty$. On the basis of this fact, we can apply the useful tuning parameter selection algorithm proposed by Wang (2010) even for the kernel normalized cut.

### 3.2. Fast rate of convergence of the normalized cut with true degree $d$

In Levrard (2015), it has been shown that, under a certain condition on the population distribution, called the margin condition, the $k$-means clustering in a Hilbert space can attain a rate of convergence of $O(1/n)$. Thus, we may expect that, under similar conditions on the population distribution, the Ncut can also achieve a fast rate of convergence. Here, we show the corresponding result that the kernel Ncut with the true degree function $d$ achieves a fast rate of convergence.

As in Levrard (2015), we first introduce some notations. Write $[1, M] = \{1, \ldots, M\}$. For any $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_M^*) \in \mathcal{M}$ and any bijection $\Pi : [1, M] \to [1, M]$, its permutation vector $\boldsymbol{\mu}' = (\mu_{\Pi(1)}^*, \ldots, \mu_{\Pi(M)}^*)$ is also an optimal codebook. Thus, we define a minimum set $\bar{\mathcal{M}}$ of the optimal codebook as a subset of $\mathcal{M}$ satisfying the following conditions:

- For all $\boldsymbol{\mu}^* \in \mathcal{M}$, there exists $\bar{\boldsymbol{\mu}} \in \bar{\mathcal{M}}$ such that $\{\mu_1^*, \ldots, \mu_M^*\} = \{\bar{\mu}_1, \ldots, \bar{\mu}_M\}$, and
- For all $\bar{\boldsymbol{\mu}}_1, \bar{\boldsymbol{\mu}}_2 \in \bar{\mathcal{M}}$ with $\bar{\boldsymbol{\mu}}_1 \neq \bar{\boldsymbol{\mu}}_2$, $\{\bar{\mu}_1^{(1)}, \ldots, \bar{\mu}_M^{(1)}\} \neq \{\bar{\mu}_1^{(2)}, \ldots, \bar{\mu}_M^{(2)}\}$.

Here, although $\bar{\mathcal{M}}$ is not uniquely determined, the cardinality of $\bar{\mathcal{M}}$ is the same. For a codebook $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)$, the Voronoi cell generated by $\mu_m$ is defined as

$$\begin{aligned}V_m(\boldsymbol{\mu}) = \{x \in \mathcal{X} \mid \forall j \neq m;\ &\|\psi_h(x) - \mu_m\|_{\mathcal{H}_h} \\ &\leq \|\psi_h(x) - \mu_j\|_{\mathcal{H}_h}\}.\end{aligned}$$

Moreover, a Voronoi partition generated by $\boldsymbol{\mu}$ is defined as a sequence $(W_1(\boldsymbol{\mu}), \ldots, W_M(\boldsymbol{\mu}))$ such that $\mathcal{X} = \bigcup_{m=1}^M W_m(\boldsymbol{\mu})$ and $\bar{W}_m(\boldsymbol{\mu}) = V_m(\boldsymbol{\mu})$ for all $m = 1, \ldots, M$, where $\bar{W}_m(\boldsymbol{\mu})$ is the closure of $W_m(\boldsymbol{\mu})$.

Before stating the main theorem, we give some fundamental properties of the kernel Ncut with the true degree function $d$. To do so, we extend the basic properties of the $k$-means described in Graf & Luschgy (2000), Graf et al. (2007) and Levrard (2015) for the weighted $k$-means.

**Proposition 5.** *Suppose that the number of elements in the support of $\mathbb{P}$ is greater than the given number of clusters, that is, $\#(\mathrm{supp}(\mathbb{P})) \geq M$. Then,*

**(i)** *for any $\boldsymbol{\mu}^* \in \mathcal{M}$, we have*

$$\forall 1 \leq m \leq M;\ \mathbb{P}(V_m(\boldsymbol{\mu}^*)) > 0.$$

**(ii)** *Accordingly, any optimal codebook $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_M^*)$ satisfies the centroid condition defined as follows:*

$$\mu_m^* = \frac{\mathbb{E}[d(X)\psi_h(X)\mathbb{1}_{W_m(\boldsymbol{\mu}^*)}(X)]}{\mathbb{E}[d(X)\mathbb{1}_{W_m(\boldsymbol{\mu}^*)}(X)]} \in \mathcal{H}_h,$$

*for $m = 1, \ldots, M$.*

**(iii)** *Moreover, for any $\boldsymbol{\mu}^* \in \mathcal{M}$ and for all $j \neq m$,*

$$\mathbb{P}\left(V_m(\boldsymbol{\mu}^*) \cap V_j(\boldsymbol{\mu}^*)\right) = 0.$$

*Proof.* See Section B.1 in the supplementary material. □

From claim (iii) in the above proposition, we have that, for any $\boldsymbol{\mu}^* \in \mathcal{M}$ and any Voronoi partition $W_m(\boldsymbol{\mu}^*)$ with $\boldsymbol{\mu}^*$,

$$\mathbb{P}(W_m(\boldsymbol{\mu}^*)) = \mathbb{P}(V_m(\boldsymbol{\mu}^*)), \text{ and}$$
$$\mathbb{P}_n(W_m(\boldsymbol{\mu}^*)) = \mathbb{P}_n(V_m(\boldsymbol{\mu}^*)) \ \ a.s.$$

Let us introduce the following key quantities:

$$B = \inf_{\boldsymbol{\mu}^* \in \mathcal{M}, m \neq j} \|\mu_m^* - \mu_j^*\|_{\mathcal{H}_h}, \text{ and}$$
$$d_{\min} = \inf_{\boldsymbol{\mu}^* \in \mathcal{M}, 1 \leq m \leq M} \mathbb{E}[d(X) \mathbb{1}_{V_m(\boldsymbol{\mu}^*)}].$$

According to the general assumption, both $B$ and $d_{\min}$ are strictly positive (see Proposition B.2 in the supplementary material.)

For any $\boldsymbol{\mu}^* \in \mathcal{M}$, we define the following set:

$$N_{\boldsymbol{\mu}^*} = \bigcup_{m=1}^{M} \bigcup_{j \neq m} V_m(\boldsymbol{\mu}^*) \cap V_j(\boldsymbol{\mu}^*).$$

For a subset $A \subseteq \mathcal{H}_h$, let $\mathcal{B}(A, D) = \bigcup_{a \in A} \mathcal{B}(a, D)$. The maximal weight of the $t$-neighborhoods of these regions $N_{\boldsymbol{\mu}^*}$ over $\mathcal{M}$ is defined as

$$q(t) = \sup_{\boldsymbol{\mu}^* \in \mathcal{M}} \mathbb{E}\Big[d(X) \mathbb{1}_{\mathcal{B}(N_{\boldsymbol{\mu}^*}, t)}(X)\Big],$$

where $X \sim \mathbb{P}$.

In analogy with Levrard (2015), we define the following margin condition with radius $r_0 > 0$: for all $0 \leq t \leq r_0$,

$$q(t) \leq \frac{B d_{\min}}{128 D^2} t. \tag{3}$$

If the weight function $d(x) = 1$ for all $x \in \mathcal{X}$, this condition is equivalent to the margin condition for $k$-means introduced in Levrard (2015). Under the margin condition, we can show identifiability, that is, there exists $\epsilon > 0$ such that

$$\inf_{\boldsymbol{\mu} \in \tilde{\mathcal{M}} \cap \mathcal{M}^c} \ell(\boldsymbol{\mu}, \boldsymbol{\mu}^*) = \epsilon > 0,$$

where $\tilde{\mathcal{M}}$ is the set of local minimizers of $\mathrm{WKKM}_h(\cdot \mid \mathbb{P})$. A distribution $\mathbb{P}$ is called $\epsilon$-separated if the above equation holds with some $\epsilon$. The proof of this claim can be founded in the supplementary material.

The kernel Ncut with the true degree function $d$ is equivalent to the weighted kernel $k$-means with the true degree function $d$, whose risk function is defined by $\mathrm{WKKM}_h(\cdot \mid \mathbb{P}_n)$. Based on this fact and the similar approach to that of Levrard (2015), we can obtain the fast rate of convergence of the kernel Ncut with the true degree function $d$.

**Theorem 6.** *Suppose that $M \geq 2$, and that $\mathbb{P}$ satisfies the margin condition (3) with radius $r_0$. Let $\kappa_0$ be given by*

$$\kappa_0 = 4MD^2 \left( \frac{1}{\epsilon} \vee \frac{64 D^2}{d_{\min} B^2 r_0^2} \right).$$

*Let $\bar{\boldsymbol{\mu}}_n$ be a minimizer of the empirical risk $\mathrm{WKKM}_h(\cdot \mid \mathbb{P}_n)$. Then, with probability at least $1 - \exp(-x)$,*

$$\ell(\bar{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) \leq C_0 \sigma^2 \kappa_0 \frac{D^2 \{M + \log(|\bar{\mathcal{M}}|)\}}{n} + (9 c_U \kappa_0 + 4) \frac{16 D^2 c_U}{n} x,$$

*where $\boldsymbol{\mu}^* \in \mathcal{M}$, $C_0$ is an universal constant, and $\sigma^2 = \mathbb{E}[d^2(X)]$. Similarly, for $\hat{\boldsymbol{\mu}}_n$ described in Section 3.1, with probability at least $1 - \exp(-x)$,*

$$\ell(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}^*) \leq C_0 \sigma^2 \kappa_0 \frac{D^2 \{M + \log(|\bar{\mathcal{M}}|)\}}{n} + (9 c_U \kappa_0 + 4) \frac{16 D^2 c_U}{n} x + 4D^2 \|\hat{d}_n - d\|_\infty.$$

*Proof.* See Section C in the supplementary material. □

From the last part of the above theorem, owing to the estimation error $\|\hat{d}_n - d\|_\infty$ of the degree function, it seems to be difficult to obtain a fast rate of convergence for the fully empirical Ncut.

## 4. Population-level relationship between Ncut and the weighted kernel $k$-means

From the results in the previous section, we know that the empirical kernelized Ncut converges to the population-level weighted $k$-means clustering on a certain RKHS. However, it is difficult to understand the meaning of this weighted $k$-means clustering on $\mathcal{H}_h$ from the viewpoint of the original data space $\mathcal{X}$ or the population distribution $\mathbb{P}$ on $\mathcal{X}$. Here, we provide an interpretable meaning for the limit of the empirical Ncut. For a partition $\mathscr{P}_M = \{W_1, \ldots, W_M\}$ of the data space $\mathcal{X}$, let

$$\mathrm{WKKM}_h(\boldsymbol{\mu} \mid \mathscr{P}_M, \mathbb{P}) := \sum_{m=1}^{M} \int_{W_m} d(x) \|\psi_h(x) - \mu_m\|_{\mathcal{H}_h}^2 \mathbb{P}(dx).$$

If $\mathbb{P}(W_m) > 0$, the optimal codebook $\boldsymbol{\mu}(\mathscr{P}_M) = (\mu_1(W_1), \ldots, \mu_M(W_M))$ of $\mathrm{WKKM}_h(\boldsymbol{\mu} | \mathscr{P}_M, \mathbb{P})$ is given by

$$\mu_m(W_m) = \frac{\int_{W_m} d(y) \psi_h(y) \mathbb{P}(dy)}{\int_{W_m} d(y) \mathbb{P}(dy)}.$$

For $\mathbb{P}(W_m) = 0$, we choose an element $\mu_m(W_m) \in W_m$. According to the Riesz representation theorem, we can easily see that $\mu_m(W_m) \in \mathcal{H}_h$ for $m = 1, \ldots, M$. Therefore,

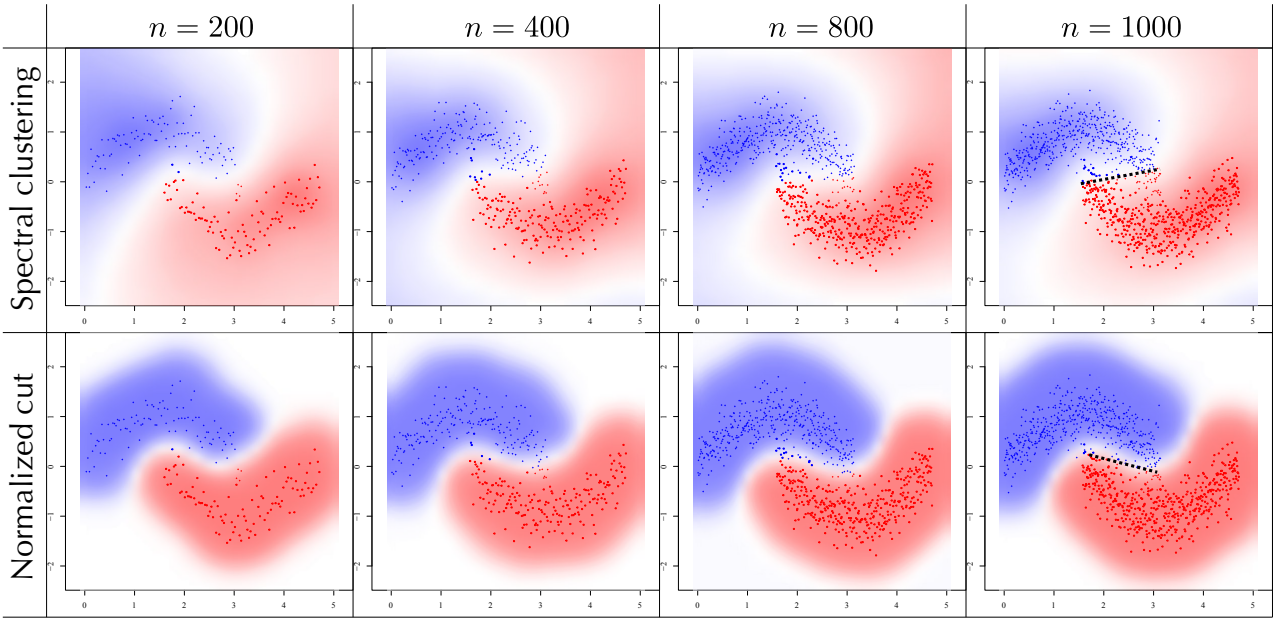


*Figure 2.* Convergence behavior of spectral clustering and normalized cut with the same tuning parameter.

we can redefine the weighted kernel $k$-means as the following optimization problem for a partition $\mathscr{P}_M$ of the data space $\mathcal{X}$:

$$\mathrm{WKKM}_h(\mathscr{P}_M \mid \mathbb{P}) := \sum_{m=1}^{M} \int_{W_m} d(x) \|\psi_h(x) - \mu_m(W_m)\|_{\mathcal{H}_h}^2 \mathbb{P}(dx).$$

Thus, in the weighted kernel $k$-means, obtaining an optimal codebook is the same as obtaining an optimal partition of $\mathcal{X}$.

The following theorem shows the equivalence between the population-level weighted kernel $k$-means and the population-level Ncut.

**Theorem 7.** *The minimization of the population-level weighted kernel $k$-means $\mathrm{WKKM}_h(\mathscr{P}_M \mid \mathbb{P})$ is equivalent to the minimization of the population-level Ncut defined as*

$$\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}) := \sum_{m=1}^{M} \frac{1}{d(W_m)} \int_{W_m} \int_{\mathcal{X} \setminus W_m} k(x,y)\, \mathbb{P}(dx)\mathbb{P}(dy). \quad (4)$$

*Proof.* See Section D in the supplementary material. □

From Theorem 4 and 7, the estimated partition of the empirical normalized cut $\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P}_n)$ converges to the optimal partition of the population-level normalized cut $\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P})$. Moreover, this equivalence may be helpful in obtaining the detailed properties of Ncut. In fact, Schiebinger et al. (2015) provided a detailed analysis for kernelized spectral clustering in which the indivisibility of $\mathbb{P}$ is defined as

$$\Gamma(\mathbb{P}) := \inf_{S} \frac{d(\mathcal{X}) \int_S \int_{S^c} k(x,y)\mathbb{P}(dx)\mathbb{P}(dy)}{d(S)d(S^c)},$$

where the infimum is taken over all measurable subset $S \subseteq \mathcal{X}$ and $d(A) = \int_A d(x)\mathbb{P}(dx)$ for $A \subset \mathcal{X}$. This criterion measures how difficult it is to *split* the population distribution $\mathbb{P}$ into two parts. In addition, the infimum of indivisibility can be interpreted as Cheeger's isoperimetric constant introduced in Lawler & Sokal (1988), which plays a key role in the convergence analysis of Markov chain. This is equivalent to the optimal value of the normalized cut $\mathrm{Ncut}(\mathscr{P}_2 \mid \mathbb{P})$ with $M = 2$.

## 5. Numerical experiments

From von Luxburg et al. (2008) and the results in the previous sections, both spectral clustering and the normalized cut converge to the corresponding limits. The limit of normalized cut is the optimal partition in the sense of the population-level normalized cut. Conversely, there is no corresponding objective function for spectral clustering and the optimality of the spectral clustering is not clear. In general, these limits are not the same. Through the numerical experiments, we show the essential difference between the spectral clustering and the kernel normalized cut.

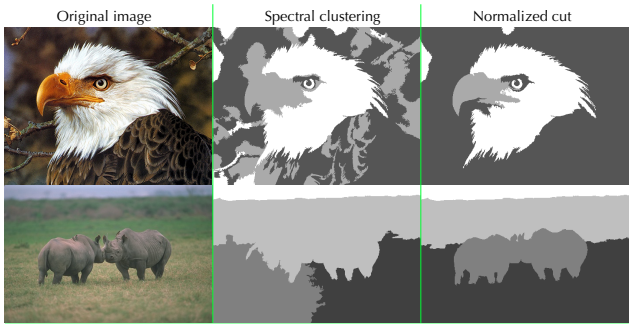


*Figure 3.* The image segmentation results of spectral clustering and normalized cut.

First, we consider the binary clustering problem for a two-moon data. For a given sample size $n$ of each cluster, we generate two dimensional data $(X_1^{(1)}, Y_1^{(1)}), \ldots, (X_n^{(1)}, Y_n^{(1)})$ and $(X_1^{(2)}, Y_1^{(2)}), \ldots, (X_n^{(2)}, Y_n^{(2)})$ as follows:

$$X_i^{(1)} \sim_{i.i.d.} U(\pi/n, \pi),\ Y_i^{(1)} = \sin(X_i^{(1)}) + 0.05 + \epsilon_i^{(1)}$$
$$X_i^{(2)} \sim_{i.i.d.} U(\pi(1/n + 1/2), \pi(1 + 1/2)),\ \text{and}$$
$$Y_i^{(2)} = \cos(X_i^{(2)}) + \epsilon_i^{(2)},$$

where $U(a, b)$ is the uniform distribution on the closed interval $[a, b]$, and $\epsilon_i^{(j)}$ is independently generated from the normal distribution with mean 0 and standard deviation 0.3, that is, $\epsilon_i^{(j)} \sim_{i.i.d.} N(0, 0.3^2)$ $(i = 1, \ldots, n;\ j = 1, 2)$. Here, we use the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\sigma^2)$ with $\sigma \approx 0.5$ as the kernel function in both spectral clustering and Ncut.

For several sample sizes $n = 200, 400, 800, 1000$, we apply both spectral clustering and Ncut with $M = 2$. Figure 2 shows the convergence behavior of these methods with the same tuning parameter. In the results of spectral clustering, the color shows the estimated value of the second smallest eigenfunction of the limit of the normalized graph Laplacian. In the results of Ncut, the color shows the value of $\|\psi_{\hat{h}_n}(x) - \tilde{\mu}_2\|_{\mathcal{H}_{\hat{h}_n}} - \|\psi_{\hat{h}_n}(x) - \tilde{\mu}_1\|_{\mathcal{H}_{\hat{h}_n}}$. From the above consistency results and this figure, it seems that these methods converge to the different limits, respectively, with increasing $n$. In Section E of the supplementary material, we describe more detailed comparisons related to the optimality in the sense of $\mathrm{Ncut}(\mathscr{P}_M \mid \mathbb{P})$.

Next, we consider image segmentation via spectral clustering and the normalized cut. From Li & Chen (2015) and the Berkeley Segmentation Data Set and Benchmarks 500 (BSDS500) in Arbelaez et al. (2011), we selected the two natural images shown in Figure 3. Since the number of pixels of each image is greater than $150,000$, we employed the fast approximation method proposed by Yan et al. (2009) to reduce the computational cost. More precisely, we first partitioned the image into a few hundred superpixels by Li & Chen (2015). For the representive points of the obtained superpixels, we applied both spectral clustering (Ng et al., 2002) and the normalized cut with the Gaussian kernel and the same tuning parameter $\sigma$. From the clustering result of the superpixels, we obtained the segmentation of the original image.

Figure 3 shows the segmentations of spectral clustering and the normalized cut. Even though we use the same kernel function and the same tuning parameter in both methods, we can see the clear difference between the results of these two methods. For the Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/(2\sigma^2))$, the tuning parameter $\sigma$ is related to the strength of the nonlinearity. In both methods, a small value of $\sigma$ induces a complicated partition, and a large value of $\sigma$ induces a result similar to $k$-means clustering. Empirically, in spectral clustering, we need to set a smaller value of $\sigma$ to obtain a similar partition to that of the Ncut. More detailed comparisons are provided in Section E of the supplementary material.

## 6. Conclusions

In this paper, we provided detailed theoretical properties of the kernel normalized cut. Whereas Dhillon et al. (2007) discovered that Ncut and the ratio cut are equivalent to the weighted kernel $k$-means methods in the sense of the optimization problem, they did not introduce a certain RKHS on which we perform the weighted kernel $k$-means. First, we showed that, under very general conditions, the codebook estimated by the empirical Ncut with the kernel function $k$ converges to the optimal solution of the population-level weighted $k$-means clustering on the RKHS $\mathcal{H}_h$ with kernel $h(x, y) = k(x, y)/(d(x)d(y))$ as the sample size goes to infinity. It is worth noting that, for the ratio cut, there is no corresponding RKHS in the large sample limit. Thus, we cannot get a similar consistency result for the ratio cut using the same framework. This difference may be related to the superiority of normalized spectral clustering compared to the unnormalized spectral clustering, which is shown in von Luxburg et al. (2008). Moreover, we proved that, under a mild condition on the population distribution, the empirical Ncut with the true degree has a fast rate of convergence. As a second contribution, we proved the equivalence between the population-level weighted kernel $k$-means and the population-level Ncut. The minimum value of the objective function of the population-level Ncut is the indivisibility of $\mathbb{P}$, introduced in Schiebinger et al. (2015). In the future work, based on these facts, we will study the performance of the Ncut in recovering the hidden cluster labels, which is parallel with the results in Schiebinger et al. (2015).

## Acknowledgements

This research was supported in part by JSPS KAKENHI Grant (16K16024 to YT, 17K12648 to MY), RIKEN Engineering Network (to YT), and a collaboration grant (to YT) from Organ Technologies Inc.

## References

Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:898–916, 2011.

Arias-Castro, E., Pelletier, B., and Pudlo, P. The normalized graph cut and cheeger constant: from discrete to continuous. *Advances in Applied Probability*, 44:907–937., 2012.

Ben-David, S., von Luxburg, U., and Pál, D. A sober look on clustering stability. In *Proceedings of the 19th Annual Conference on Learning Theory*, pp. 5–19, 2006.

Cao, Y. and Chen, A. D.-R. Consistency of regularized spectral clustering. *Applied and Computational Harmonic Analysis*, 30:319–336, 2011.

Davis, E. and Sethuraman, S. Consistency of modularity clustering on random geometric graphs. *arXiv*, 2016.

Dhillon, I., Guan, Y., and Kulis, B. Weighted graph cuts without eigenvectors: A multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29: 1944–1957, 2007.

Donath, W. E. and Hoffman, A. J. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17:420–425, 1973.

Fideler, M. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.

Gracía Trillos, N. and Slepčev, D. A variational approach to the consistency of spectral clustering. *to appear in Applied and Computational Harmonic Analysis*, 2016.

Gracía Trillos, N., Slepčev, D., von Brechet, J., Laurent, T., and Bresson, X. Consistency of cheeger and ratio graph cuts. *Journal of Machine Learning Research*, 17:1–46, 2016.

Graf, S. and Luschgy, H. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, 2000.

Graf, S., Luschgy, H., and Pagès, G. Optimal quantizers for radon random vectors in a banach space. *Journal of Approximation Theory*, 144:27–53, 2007.

Lawler, G. F. and Sokal, A. D. Bounds on the $l^2$ spectrum for markov chains and markov processes: A generalization of cheeger's inequality. *Transactions of the American Mathematical Society*, 309:557–580, 1988.

Levrard, C. Nonasymptotic bounds for vector quantization in hilbert spaces. *Annals of Statistics*, 43:592–619, 2015.

Li, Z. and Chen, J. Superpixel segmentation using linear spectral clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1356–1363. IEEE, 2015.

Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: analysis and an algorithm. In *Proceedings of Advances in Neural Information Processing Systems 14*, pp. 849–856, 2002.

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

Rosasco, L., Belkin, M., and Vito, E. D. On learning with integral operators. *Journal of Machine Learning Research*, 11:905–934, 2010.

Schiebinger, G., Wainwright, M. J., and Yu, B. The geometry of kernelized spectral clustering. *Annals of Statistics*, 43:818–846, 2015.

Shi, J. and Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

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

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

Wang, J. Consistent selection of the number of clusters via cross validation. *Biometrika*, 97:893–904, 2010.

Yan, D., Huang, L., and Jordan, M. I. Fast approximate spectral clustering. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 907–916, 2009.