

CONSISTENCY OF CHEEGER AND RATIO GRAPH CUTS

NICOLÁS GARCÍA TRILLOS¹, DEJAN SLEPČEV¹, JAMES VON BRECHT², THOMAS LAURENT³, XAVIER BRESSON⁴.

ABSTRACT. This paper establishes the consistency of a family of graph-cut-based algorithms for clustering of data clouds. We consider point clouds obtained as samples of a ground-truth measure. We investigate approaches to clustering based on minimizing objective functionals defined on proximity graphs of the given sample. Our focus is on functionals based on graph cuts like the Cheeger and ratio cuts. We show that minimizers of these cuts converge as the sample size increases to a minimizer of a corresponding continuum cut (which partitions the ground truth measure). Moreover, we obtain sharp conditions on how the connectivity radius can be scaled with respect to the number of sample points for the consistency to hold. We provide results for two-way and for multiway cuts. Furthermore we provide numerical experiments that illustrate the results and explore the optimality of scaling in dimension two.

1. INTRODUCTION

Partitioning data clouds in meaningful clusters is one of the fundamental tasks in data analysis and machine learning. A large class of the approaches, relevant to high-dimensional data, relies on creating a graph out of the data cloud by connecting nearby points. This allows one to leverage the geometry of the data set and obtain high quality clustering. Many of the graph-clustering approaches are based on optimizing an objective function which measures the quality of the partition. The basic desire to obtain clusters which are well separated leads to the introduction of objective functionals which penalize the size of cuts between clusters. The desire to have clusters of meaningful size and for approaches to be robust to outliers leads to the introduction of "balance" terms and objective functionals such as Cheeger cut (closely related to edge expansion) [5, 11, 12, 28, 38], ratio cut [23, 27, 40, 43], normalized cut [4, 34, 40], and conductance (sparsest cut) [5, 28, 36]. Such functionals can be extended to treat multiclass partitioning [13, 45]. The balanced cuts above have been widely studied theoretically and used computationally. The algorithms of [2, 36, 37] utilize local clustering algorithms to compute balanced cuts of large graphs. Total variation based algorithms [12, 13, 26, 27, 38] are also used to optimize either the conductance or the edge expansion of a graph. Closely related are the spectral approaches to clustering [34, 40] which can be seen as a relaxation of the normalized cuts.

In this paper we consider data clouds, $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, which have been obtained as i.i.d. samples of a measure ν with density ρ on a bounded domain D . The measure ν represents the ground truth that X_n is a sample of. In the large sample limit, $n \rightarrow \infty$, clustering methods should exhibit *consistency*. That is, the clustering of the data sample X_n should converge as $n \rightarrow \infty$ toward a specific clustering of the underlying sample domain. In this paper we characterize in a precise manner when and how the minimizers of a ratio and Cheeger graph cuts converge towards a suitable partition of the domain. We define the discrete and continuum objective functionals considered in Subsections 1.1 and 1.2 respectively, and informally state our result in Subsection 1.3.

Date: November 24, 2014.

1991 Mathematics Subject Classification. 62H30, 62G20, 49J55, 91C20, 68R10, 60D05.

Key words and phrases. data clustering, balanced cut, consistency, Gamma convergence, graph partitioning.

An important consideration when investigating consistency of algorithms is how the graphs on X_n are constructed. In simple terms, when building a graph on X_n one sets a length scale ε_n such that edges between vertices in X_n are given significant weights if the distance of points they connect is ε_n or less. In some way this sets the length scale over which the geometric information is averaged when setting up the graph. Taking smaller ε_n is desirable because it is computationally less expensive and gives better resolution, but there is a price. Taking ε_n small increases the error due to randomness and in fact if ε_n is too small the resulting graph may not represent the geometry of D well and consequently the discrete graph cut may be very far from the desired one. In our work we determine precisely how small ε_n can be taken for the consistency to hold. We obtain consistency results both for two-way and multiway cuts.

To prove our results we use the variational notion of convergence known as the Γ -convergence. It is one of the standard tools of modern applied analysis to consider a limit of a family of variational problems [10, 16]. In the recent work [18], this notion was developed in the random discrete setting designed for the study of consistency of minimization problems on random point clouds. In particular the proof of Γ -convergence of total variation on graphs proved there provides the technical backbone of this paper. The approach we take is general and flexible and we believe suitable for the study of many problems involving large sample limits of minimization problems on graphs.

Background on consistency of clustering algorithms and related problems. Consistency of clustering algorithms has been considered for a number of approaches. Pollard [33] has proved the consistency of k -means clustering. Consistency for a class of single linkage clustering algorithms was shown by Hartigan [24]. Arias-Castro and Pelletier have proved the consistency of maximum variance unfolding [3]. Consistency of spectral clustering was rigorously considered by von Luxburg, Belkin, and Bousquet [41, 42]. These works show the convergence of all eigenfunctions of the graph laplacian for fixed length scale $\varepsilon_n = \varepsilon$ which results in the limiting (as $n \rightarrow \infty$) continuum problem being a nonlocal one. Belkin and Niyogi [7] consider the spectral problem (Laplacian eigenmaps) and show that there exists a sequence $\varepsilon_n \rightarrow 0$ such that in the limit the (manifold) Laplacian is recovered, however no rate at which ε_n can go to zero is provided. Consistency of normalized cuts was considered by Arias-Castro, Peletier, and Pudlo [4] who provide a rate on $\varepsilon_n \rightarrow 0$ under which the minimizers of the discrete cut functionals minimized over a specific family of subsets of X_n converge to the continuum Cheeger set. Our work improves on [4] in several ways. We minimize the discrete functionals over all discrete partitions on X_n as it is considered in practice and prove the result for the optimal, in terms of scaling, range of rates at which $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$.

There are also a number of works which investigate how well the discrete functionals approximate the continuum ones for a particular function. Among them are works by Belkin and Niyogi [8], Giné and Koltchinskii [20], Hein, Audibert, von Luxburg [25], Singer [35] and Ting, Huang, and Jordan [39]. Maier, von Luxburg and Hein [30] considered pointwise convergence for Cheeger and normalized cuts, both for the geometric and kNN graphs and obtained a range of scalings of graph construction on n for the convergence to hold. While these results are quite valuable, we point out that they do not imply that the minimizers of discrete objective functionals are close to minimizers of continuum functionals.

1.1. Graph partitioning. The balanced cut objective functionals we consider are relevant to general graphs (not just ones obtained from point clouds). We introduce them here.

Given a weighted graph $\mathcal{G} = (X, W)$ with the vertex set $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ and the weight matrix $W = \{w_{ij}\}_{1 \leq i, j \leq n}$, the balanced graph cut problems we consider take the form

$$(1.1) \quad \text{Minimize} \quad \frac{\text{Cut}(Y, Y^c)}{\text{Bal}(Y, Y^c)} := \frac{\sum_{\mathbf{x}_i \in Y} \sum_{\mathbf{x}_j \in Y^c} w_{ij}}{\text{Bal}(Y, Y^c)} \quad \text{over all nonempty } Y \subsetneq X.$$

That is, we consider the class of problems with $\text{Cut}(Y, Y^c)$ as the numerator together with different balance terms. For $Y \subset X$ let $|Y|$ be the ratio between the number of vertices in Y and the number of vertices in X . Well-known balance terms include

$$(1.2) \quad \text{Bal}_R(Y, Y^c) = 2|Y||Y^c| \quad \text{and} \quad \text{Bal}_C(Y, Y^c) = \min(|Y|, |Y^c|),$$

which correspond to Ratio Cut [23, 27, 40, 43] and Cheeger Cut [5, 14, 15, 28] respectively ¹. A variety of other balance terms have appeared in the literature in the context of two-class and multiclass clustering [11, 27]. We refer to a pair $\{Y, Y^c\}$ that solves (1.1) as an *optimal balanced cut of the graph*. Note that a given graph $\mathcal{G} = (X, W)$ may have several optimal balanced cuts (although generically the optimal cut is unique).

We are also interested in multiclass balance cuts. Specifically, in order to partition the set X into $K \geq 3$ clusters, we consider the following ratio cut functional:

$$(1.3) \quad \underset{(Y_1, \dots, Y_K)}{\text{Minimize}} \quad \sum_{k=1}^K \frac{\text{Cut}(Y_k, Y_k^c)}{|Y_k|}, \quad Y_k \cap Y_s = \emptyset \quad \text{if } r \neq s, \quad \bigcup_{k=1}^K Y_k = X.$$

1.2. Continuum partitioning. Given a bounded and connected open domain $D \subset \mathbb{R}^d$ and a probability measure ν on D , with positive density $\rho > 0$, we define the class of balanced domain cut problems in an analogous way. A balanced domain-cut problem takes the form

$$(1.4) \quad \underset{A \subset D}{\text{Minimize}} \quad \frac{\text{Cut}_\rho(A, A^c)}{\text{Bal}_\rho(A, A^c)}, \quad A \subset D \text{ with } 0 < \nu(A) < 1.$$

where $A^c = D \setminus A$. Just as the graph cut term $\text{Cut}(Y, Y^c)$ in (1.1) provides a weighted (by W) measure of the boundary between Y and Y^c , the cut term $\text{Cut}_\rho(A, A^c)$ for a domain denotes a ρ^2 -weighted area of the boundary between the sets A and A^c . If $\partial_D A := \partial A \cap D$ (the boundary between A and A^c) is a smooth curve (in 2d), surface (in 3d) or manifold (in 4d+) then we define

$$(1.5) \quad \text{Cut}_\rho(A, A^c) := \int_{\partial_D A} \rho^2(x) \, dS(x).$$

For our results and analysis we need the notion of continuum cut which is defined for sets with less regular boundary. We present the required notions of geometric measure theory and the rigorous and mathematically precise formulation of problem (1.4) in Subsection 3.1.

If $\rho(x) = 1$ then $\text{Cut}_\rho(A, A^c)$ simply corresponds to arc-length (in 2d) or surface area (in 3d). In the general case, the presence of $\rho^2(x)$ in (1.5) indicates that the regions of low density are easier to cut, so ∂A has a tendency to pass through regions in D of low density. As in the graph case, we consider balance terms

$$(1.6) \quad \text{Bal}_\rho(A, A^c) = 2|A||A^c| \quad \text{and} \quad \text{Bal}_\rho(A, A^c) = \min(|A|, |A^c|),$$

which correspond to weighted continuous equivalents of the Ratio Cut and the Cheeger Cut. In the continuum setting $|A|$ stands for the total ν -content of the set A , that is,

$$(1.7) \quad |A| = \nu(A) = \int_A \rho(x) \, dx.$$

We refer to a pair $\{A, A^c\}$ that solves (1.4) as an *optimal balanced cut of the domain*.

¹The factor of 2 in the definition of $\text{Bal}_R(Y, Y^c)$ is introduced to simplify the computations in the remainder. We remark that when using Bal_R , problem (1.1) is equivalent to the usual ratio cut problem.

The continuum equivalent of the multiway cut problem (1.3) reads

$$(1.8) \quad \underset{(A_1, \dots, A_R)}{\text{Minimize}} \quad \sum_{r=1}^R \frac{\text{Cut}_\rho(A_r, A_r^c)}{|A_r|}, \quad A_r \cap A_s = \emptyset \quad \text{if } r \neq s, \quad \bigcup_{r=1}^R A_r = D.$$

1.3. Consistency of partitioning of data clouds. We consider the sample $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ consisting of i.i.d. random points drawn from an underlying ground-truth measure ν . We assume that D is a bounded, open set with Lipschitz boundary. Furthermore we assume that ν has continuous density ρ and that $\lambda \leq \rho \leq \Lambda$ on D .

To extract the desired information about the point cloud one builds a graph by connecting the nearby points. More precisely consider a kernel $\eta : \mathbb{R}^d \rightarrow [0, \infty)$ to be radially symmetric, radially decreasing, and decaying to zero sufficiently fast. We introduce a parameter ε which basically describes over which length scale the data points are connected. We assign for $i, j \in \{1, \dots, n\}$ the weights by

$$(1.9) \quad w_{i,j} = \eta \left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|}{\varepsilon} \right).$$

As more data points are available one takes smaller ε to obtain increased resolution. That is, one sets the length scale ε_n based on the number of available data points. We investigate under what scaling of ε_n on n the optimal balanced cuts (that is minimizers of (1.1)) of the graph $\mathcal{G}_n = (X_n, W_n)$ converge towards optimal balanced cuts in the continuum setting (minimizers of (1.4)). On Figure 1, we illustrate partitioning a data cloud sampled from the uniform distribution for the given domain D .

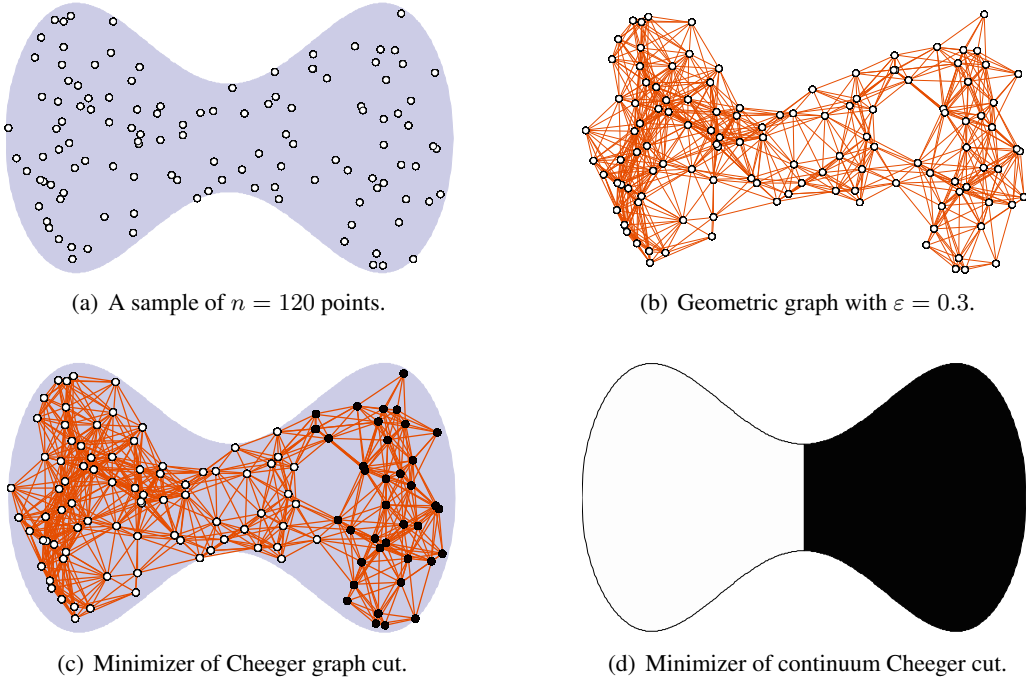


FIGURE 1. Given the sample of Figure (a), graph is constructed using $\eta(z) = \mathbf{1}_{\{|z| \leq 1\}}$ and $\varepsilon = 0.3$, as illustrated on Figure (b). On Figure (c) we present the solution to the Cheeger graph-cut problem obtained using [12]. A solution to the continuum Cheeger-cut problem is illustrated in Figure (d).

Informal statement of (a part of) the main results. Consider $d \geq 2$ and assume the continuum balanced cut (1.4) has a unique minimizer $\{A, A^c\}$. Consider $\varepsilon_n > 0$ such that $\lim_{n \rightarrow \infty} \varepsilon_n = 0$ and

$$\lim_{n \rightarrow \infty} \frac{(\log n)^{p_d}}{n^{1/d}} \frac{1}{\varepsilon_n} = 0,$$

where $p_d = 1/d$ for $d \geq 3$ and $p_2 = 3/4$. Then almost surely the minimizers, $\{Y_n, Y_n^c\}$, of the balanced cut (1.1) of the graph \mathcal{G}_n , converge to $\{A, A^c\}$. Moreover, after appropriate rescaling, almost surely the minimum of problem (1.1) converges to the minimum of (1.4). The result also holds for multiway cuts. That is the minimizers of (1.3) converge towards minimizers of (1.8).

Let us make the notion of convergence of discrete partitions $\{Y_n, Y_n^c\}$ to continuum partitions $\{A, A^c\}$ precise. Let $\mathbf{1}_{Y_n} : X_n \rightarrow \{0, 1\}$ be the characteristic function of Y_n on the set X_n . Let $\mathbf{p}_{i,1} = (\mathbf{x}_i, \mathbf{1}_{Y_n}(\mathbf{x}_i))$ and $\mathbf{p}_{i,2} = (\mathbf{x}_i, \mathbf{1}_{Y_n^c}(\mathbf{x}_i))$ for $i = 1, \dots, n$ be the points on the graphs of $\mathbf{1}_{Y_n}$ and $\mathbf{1}_{Y_n^c}$ respectively.

Consider the probability measures on the graphs of $\mathbf{1}_{Y_n}$ and $\mathbf{1}_{Y_n^c}$: that is let $\gamma_{n,1} := \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{p}_{i,1}}$ and $\gamma_{n,2} := \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{p}_{i,2}}$. Let γ be the push-forward of the measure ν to the graph of $\mathbf{1}_A$, that is $\gamma = (Id \times \mathbf{1}_A)_\# \nu = \nu_{\perp A} \times \delta_1 + \nu_{\perp A^c} \times \delta_0$ where Id is the identity mapping from \mathbb{R}^d to \mathbb{R}^d and $\nu_{\perp A}$ is the restriction of the measure ν to the set A . We say that $\{Y_n, Y_n^c\}$ converge towards $\{A, A^c\}$ as $n \rightarrow \infty$ if there is a sequence of indices $I : \mathbb{N} \rightarrow \{1, 2\}$ such that

$$(1.10) \quad \gamma_{n,I(n)} \xrightarrow{w} \gamma \quad \text{as } n \rightarrow \infty,$$

which is to be read as $\gamma_{n,I(n)}$ converges weakly to γ (see [17] for the definition of weak convergence of probability measures.) In other words the convergence of discrete towards continuum partitions is defined as the weak convergence of graphs, considered as probability measures.

In Section 2 we discuss this topology in more detail and present a more general and conceptually clearer picture. In particular we point out that the weak convergence of measures on the space of graphs of functions is stronger than it may look and actually corresponds to L^1 convergence of functions.

Remark 1 (Optimality of scaling of ε_n for $d \geq 3$). If $d \geq 3$ then the rate presented in the statement above is sharp in terms of scaling. Namely for $D = (0, 1)^d$, ν being the Lebesgue measure on D and η compactly supported, it is known from graph theory (see [21, 22, 31]) that there exists a constant $c > 0$ such that if $\varepsilon_n < c \frac{(\log n)^{1/d}}{n^{1/d}}$ then the weighted graph associated to (X_n, W_n) is disconnected with high probability. The resulting optimal discrete cuts have zero energy, but may be very far from the optimal continuum cuts. While the above example demonstrates the optimality of our results, we caution that there may be settings relevant to machine learning in which the convergence of minimizers of appropriate functionals involving perimeter may hold even when $\frac{1}{n^{1/d}} \ll \varepsilon_n < c \frac{(\log n)^{1/d}}{n^{1/d}}$.

Remark 2. In case $d = 2$ the connectivity threshold for a random geometric graph is $\varepsilon_n = c \frac{\log(n)^{1/2}}{n^{1/2}}$, which is below the rate for which we can establish the consistency of balanced cuts. Thus, an interesting open problem is to determine if the consistency results we present in this paper are still valid when the parameter ε_n is taken below the rate $\frac{\log(n)^{3/4}}{n^{1/2}}$ we obtained the proof for, but above the connectivity rate. In particular we are interested in determining if connectivity is the determining factor in order to obtain consistency of balance graph cuts. We numerically explore this problem in Section 8.

We also remark that, despite the fact that for a general graphs the problems (1.1) and (1.3) are NP hard, in practice when the graph is obtained by sampling from a measure ν as above, such minimization problems can be effectively approached [12, 13]. In fact, by choosing an appropriate initialization, the algorithms (see [12, 13]), give very good results in clustering real-world data.

1.4. Outline. In Section 2 we introduce the notion of convergence we use to bridge between discrete and continuum partitions. It relies on some of the notions of the theory of optimal transportation which we recall. Finally we recall results on optimal min-max matching which are needed in the proof of the convergence. In Section 3 we study more carefully continuum partitioning (1.4). We introduce the notion of total variation of functions on D in Subsection 3.1 and recall some of its basic properties. It enables us to introduce, in Subsection 3.2, the general setting for problem (1.4) where desirable properties such as lower semicontinuity and existence of minimizers hold. In Section 4 we give the precise statement of the consistency result, both for the two-way cuts and the multiple-way cuts. Proving that minimizers of discrete balanced cuts converge to continuum balanced cuts relies on a notion of variational convergence known as Γ -convergence. In Section 5 we recall the definition of Γ convergence and its basic properties. In Subsection 5.1 we recall the results on Γ -convergence of graph total variation which provide the backbone for our result. Section 6 contains the proof of Theorem 8 and Section 7 the proof of Theorem 10. Finally, in Section 8 we present numerical experiments which illustrate our results and we also investigate the issues related to Remark 2.

2. FROM DISCRETE TO CONTINUUM

For the two-class case, our main result shows that a sequence of partitions $\{Y_n, Y_n^c\}$ of the sample points $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset D$ converges toward a continuum partition $\{A, A^c\}$ of the domain D . In this section we expand on the notion of convergence introduced in Subsection 1.3 to compare the discrete and continuum partitions. We give an equivalent definition for such type of convergence which turns out to be more useful for the computations in the remainder.

Associated to the partitions $\{Y_n, Y_n^c\}$ of $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ there are characteristic functions of Y_n and Y_n^c , namely $\mathbf{1}_{Y_n} : X_n \rightarrow \{0, 1\}$ and $\mathbf{1}_{Y_n^c} : X_n \rightarrow \{0, 1\}$. Let $\nu_n = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ be the empirical measures associated to X_n . Note that $\mathbf{1}_{Y_n}, \mathbf{1}_{Y_n^c} \in L^1(\nu_n)$. Likewise a continuum partition of D by measurable sets A and $A^c = D \setminus A$ can be described via the characteristic functions $\mathbf{1}_A : D \rightarrow \{0, 1\}$ and $\mathbf{1}_{A^c} : D \rightarrow \{0, 1\}$. These too can be considered as L^1 functions, but with respect to the measure ν rather than ν_n .

We compare partitions the $\{Y_n, Y_n^c\}$ and $\{A, A^c\}$ by comparing the associated characteristic functions. To do so, we need a way of comparing L^1 functions with respect to different measures. We follow the approach of [18]. We denote by $\mathcal{B}(D)$ the Borel σ -algebra on D and by $\mathcal{P}(D)$ the set of Borel probability measures on D . The set of objects of our interest is

$$TL^1(D) := \{(\mu, f) : \mu \in \mathcal{P}(D), f \in L^1(\mu)\}.$$

Note that $(\nu_n, \mathbf{1}_{Y_n})$ and $(\nu, \mathbf{1}_A)$ both belong to TL^1 . To compare functions defined with respect to different measures, say (μ, f) and (θ, g) in TL^1 , we need a way to say for which $(x, y) \in \text{supp}(\mu) \times \text{supp}(\theta)$ should we compare $f(x)$ and $g(y)$. The notion of *coupling* (or *transportation plan*) between μ and θ , provides a way to do that. A coupling between $\mu, \theta \in \mathcal{P}(D)$ is a probability measure π on the product space $D \times D$, such that the marginal on the first variable is μ and the marginal on the second variable is θ . The set of couplings $\Gamma(\mu, \theta)$ is thus

$$\Gamma(\mu, \theta) = \{\pi \in \mathcal{P}(D \times D) : (\forall U \in \mathcal{B}(D)) \pi(U \times D) = \mu(U) \text{ and } \pi(D \times U) = \theta(U)\}.$$

For (μ, f) and (θ, g) in $TL^1(D)$ we define the distance

$$(2.1) \quad d_{TL^1}((\mu, f), (\theta, g)) = \inf_{\pi \in \Gamma(\mu, \theta)} \iint_{D \times D} |x - y| + |f(x) - g(y)| d\pi(x, y).$$

This is the distance that we use to compare L^1 functions with respect to different measures. To understand it better we focus on the case that one of the measures, say μ , is absolutely continuous with

respect to the Lebesgue measure, as this case is relevant for us when passing from discrete to continuum. In this case the convergence in TL^1 space can be formulated in simpler ways using transportation maps instead of couplings to match the measures. Given a Borel map $T : D \rightarrow D$ and $\mu \in \mathcal{P}(D)$ the *push-forward* of μ by T , denoted by $T_{\#}\mu \in \mathcal{P}(D)$ is given by:

$$T_{\#}\mu(A) := \mu(T^{-1}(A)), \quad A \in \mathfrak{B}(D).$$

A Borel map $T : D \rightarrow D$ is a *transportation map* between the measures $\mu \in \mathcal{P}(D)$ and $\theta \in \mathcal{P}(D)$ if $\theta = T_{\#}\mu$. Associated to a transportation map T , there is a plan $\pi_T \in \Gamma(\mu, \theta)$ given by $\pi_T := (\text{Id} \times T)_{\#}\mu$, where $(\text{Id} \times T)(x) = (x, T(x))$.

We note that if $\theta = T_{\#}\mu$ then the following change of variables formula holds for any $f \in L^1(\theta)$

$$(2.2) \quad \int_D f(y) d\theta(y) = \int_D f(T(x)) d\mu(x).$$

In order to give the desired interpretation of convergence in TL^1 we also need the notion of a stagnating sequence of transportation maps. A sequence $\{T_n\}_{n \in \mathbb{N}}$ of transportation maps between μ and $\{\mu_n\}_{n \in \mathbb{N}}$ (i.e. $T_{n\#}\mu = \mu_n$) is *stagnating* if

$$(2.3) \quad \int_D |x - T_n(x)| d\mu(x) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

This notion is relevant to our considerations since for the measure ν and its empirical measures ν_n there exists (with probability one) a sequence of stagnating transportation maps $T_{n\#}\nu = \nu_n$. The idea is that as $n \rightarrow \infty$ the mass from ν needs to be moved only very little to be matched with the mass of ν_n . We make this precise in Proposition 4

We now provide the desired interpretation of the convergence in TL^1 , which is a part of Proposition 3.12 in [18].

Proposition 3. *Consider a measure $\mu \in \mathcal{P}(D)$ which is absolutely continuous with respect to the Lebesgue measure. Let $(\mu, f) \in TL^1(D)$ and let $\{(\mu_n, f_n)\}_{n \in \mathbb{N}}$ be a sequence in $TL^1(D)$. The following statements are equivalent:*

- (i) $(\mu_n, f_n) \xrightarrow{TL^1} (\mu, f)$ as $n \rightarrow \infty$.
 - (ii) $\mu_n \xrightarrow{w} \mu$ and there exists a stagnating sequence of transportation maps $T_{n\#}\mu = \mu_n$ such that:
- $$(2.4) \quad \int_D |f(x) - f_n(T_n(x))| d\mu(x) \rightarrow 0, \text{ as } n \rightarrow \infty.$$
- (iii) $\mu_n \xrightarrow{w} \mu$ and for any stagnating sequence of transportation maps $T_{n\#}\mu = \mu_n$ convergence (2.4) holds.

The previous proposition implies that in order to show that (μ_n, f_n) converges to (μ, f) in the TL^1 -sense, it is enough to find a sequence of stagnating transportation maps $T_{n\#}\mu = \mu_n$ and then show the L^1 convergence of $f_n \circ T_n$ to f . An important feature of Proposition 3 is that there is complete freedom on what sequence of transportation maps $\{T_n\}_{n \in \mathbb{N}}$ to take, as long as it is stagnating. In particular this shows that if $\mu_n = \mu$ for all n then the convergence in TL^1 is equivalent to convergence in L^1 .

To apply the above to our setting we need a stagnating sequence of transportation maps between ν and $\{\nu_n\}_{n \in \mathbb{N}}$. The results on optimal transportation provide such a sequence with precise information on the rate at which (2.3) occurs. For some of our considerations it is useful to have the control of $x - T_n(x)$ in the stronger L^∞ -norm, rather than in the L^1 -norm. The following result of [19] provides such transportation maps with optimal scaling of the norm on n .

Proposition 4. *Let D be an open, connected and bounded subset of \mathbb{R}^d which has Lipschitz boundary. Let ν be a probability measure on D with density ρ which is bounded from below and from above by positive constants. Let $\mathbf{x}_1, \dots, \mathbf{x}_n, \dots$ be a sequence of independent random points distributed on D according to measure ν and let $\nu_n = \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}$. Then there is a constant $C > 0$ such that with probability one there exists a sequence of transportation maps $\{T_n\}_{n \in \mathbb{N}}$ from ν to ν_n ($T_n \# \nu = \nu_n$) and such that:*

$$(2.5) \quad \limsup_{n \rightarrow \infty} \frac{n^{1/d} \|Id - T_n\|_\infty}{(\log n)^{p_d}} \leq C,$$

where the power p_d is equal to $1/d$ if $d \geq 3$ and equal to $3/4$ if $d = 2$.

Having defined the TL^1 -convergence for functions, we turn to defining the TL^1 -convergence for partitions. When formalizing a notion of convergence for sequences of partitions $\{Y_1^n, \dots, Y_R^n\}$, we need to address the inherent ambiguity that arises from the fact that both $\{Y_1^n, \dots, Y_R^n\}$ and $\{Y_{P(1)}^n, \dots, Y_{P(R)}^n\}$ refer to the same partition for any permutation P on the set $\{1, \dots, R\}$. Having the previous observation in mind, the convergence of partitions is defined in a natural way.

Definition 5. *The sequence $\{Y_1^n, \dots, Y_R^n\}_{n \in \mathbb{N}}$, where $\{Y_1^n, \dots, Y_R^n\}$ is a partition of X_n , converges in the TL^1 -sense to the partition $\{A_1, \dots, A_R\}$ of D , if there exists a sequence of permutations $\{P_n\}_{n \in \mathbb{N}}$ of the set $\{1, \dots, R\}$, such that for every $r \in \{1, \dots, R\}$,*

$$\left(\nu_n, \mathbf{1}_{Y_{P_n(r)}^n} \right) \xrightarrow{TL^1} (\nu, \mathbf{1}_{A_r}) \quad \text{as } n \rightarrow \infty.$$

We note that the definition above is equivalent to the definition in (1.10) which we gave in Subsection (1.3) when discussing the main result. The equivalence follows from the fact that the TL^1 metric (2.1) can be seen as the distance between the graphs of functions, considered as measures. Namely given $(\mu, f), (\theta, g) \in TL^1(D)$, let $\Gamma_f = (\text{Id} \times f) \# \mu$ and $\Gamma_g = (\text{Id} \times g) \# \theta$ be the measures representing the graphs. Consider $d(\Gamma_f, \Gamma_g) := d_{TL^1}((\mu, f), (\theta, g))$. Proposition 3.3 in [18] implies that this distance metrizes the weak convergence of measures on the family of graph measures. Therefore the convergence of partitions of Definition 5 is equivalent to one given in (1.10).

We end this section by making some remarks about why the TL^1 -metric is a suitable metric for considering consistency problems. On one hand if one considers a sequence of minimizers $\{Y_n, Y_n^c\}$ of the graph balanced cut (1.1) the topology needs to be weak enough for the sequence of minimizers to be guaranteed to converge (at least along a subsequence). Mathematically speaking the topology needs to be weak enough for the sequence to be pre-compact. On the other hand the topology has to be strong enough for one to be able to conclude that the limit of a sequence of minimizers is a minimizer of the continuum balanced cut energy. In Proposition 19 and Lemma 21 we establish that the TL^1 -metric satisfies both of the desired properties.

Finally we point out that our approach from discrete to continuum can be interpreted as an extrapolation or extension approach, as opposed to restriction viewpoint. Namely when comparing (μ_n, f_n) and (μ, f) where μ_n is discrete and μ is absolutely continuous with respect to the Lebesgue measure we end up comparing two L^1 functions with respect to the Lebesgue measure, namely $f_n \circ T_n$ and f , in (2.4). Therefore $f_n \circ T_n$ used in Proposition 3 can be seen as a continuum representative (extrapolation) of the discrete f_n . We think that this approach is more flexible and suitable for the task than the, perhaps more common, approach of comparing the discrete and continuum by restricting the continuum object to the discrete setting (this would correspond to considering $f|_{\text{supp}(\mu_n)}$ and comparing it to f_n).

3. CONTINUUM PARTITIONING: RIGOROUS SETTING

We first recall the general notion of (weighted) total variation and some notions of analysis and geometric measure theory.

3.1. Total Variation. Let D be an open and bounded domain in \mathbb{R}^d with Lipschitz boundary and let $\rho : D \rightarrow (0, \infty)$ be a continuous density function. We let ν be the measure with density ρ . We assume that ρ is bounded above and below by positive constants, that is, $\lambda \leq \rho \leq \Lambda$ on D for some $\Lambda \geq \lambda > 0$. If needed, we consider an extension of ρ to the whole \mathbb{R}^d by setting $\rho(x) = \lambda$ for $x \in \mathbb{R}^d \setminus D$. This extension is a lower semi-continuous function and has the same lower and upper bounds that the original ρ has.

Given a function $u \in L^1(\nu)$, we define the weighted (by weight ρ^2) total variation of u by:

$$(3.1) \quad TV(u; D) := \sup \left\{ \int_D u(x) \operatorname{div}(\Phi(x)) \, dx : \Phi(x) \in C_c^1(D; \mathbb{R}^d), \quad |\Phi(x)| \leq \rho^2(x) \right\}.$$

If u is regular enough then the weighted total variation can be written as

$$(3.2) \quad TV(u; D) = \int_D |\nabla u| \rho^2(x) \, dx.$$

Also, given that $\rho : D \rightarrow \mathbb{R}$ is continuous, if $u = \mathbf{1}_A$ is the characteristic function of a set $A \subseteq \mathbb{R}^d$ with C^1 boundary, then

$$(3.3) \quad TV(\mathbf{1}_A; D) = \int_{\partial A \cap D} \rho^2(x) \, d\mathcal{H}^{d-1}(x),$$

where \mathcal{H}^{d-1} represents the $(d-1)$ -dimensional Hausdorff measure in \mathbb{R}^d . In case ρ is a constant (ν is the uniform distribution), the functional $TV(\cdot; D)$ reduces to a multiple of the classical total variation and in particular (3.3) reduces to a multiple of the surface area of the portion of ∂A contained in D .

Since ρ is bounded above and below by positive constants, a function $u \in L^1(\nu)$ has finite weighted total variation if and only if it has finite classical total variation. Therefore, if $u \in L^1(\nu)$ with $TV(u; D) < \infty$, then u is a BV function and hence it has a distributional derivative Du which is a Radon measure (see Chapter 13 in [29]). We denote by $|Du|$ the total variation of the measure Du and denote by $|Du|_{\rho^2}$ the measure determined by

$$(3.4) \quad d|Du|_{\rho^2} = \rho^2(x) d|Du|.$$

By Theorem 4.1 in [6]

$$(3.5) \quad TV(u; D) = |Du|_{\rho^2}(D) = \int_D \rho^2(x) \, d|Du|(x).$$

A simple consequence of the definition of the weighted TV is its lower semicontinuity with respect to L^1 -convergence. More precisely, if $u_k \xrightarrow{L^1(\nu)} u$ then

$$(3.6) \quad TV(u; D) \leq \liminf_{k \rightarrow \infty} TV(u_k; D).$$

Finally, for $u \in BV(D)$, the co-area formula

$$TV(u; D) = \int_{\mathbb{R}} TV(\mathbf{1}_{\{u>t\}}; D) \, dt,$$

relates the weighted total variation of u with the weighted total variation of its level sets. A proof of this formula can be found in [9]. For a proof of the formula in case ρ is constant see [29].

In the remainder of the paper, we write $TV(u)$ instead of $TV(u; D)$ when the context is clear.

3.2. Continuum partitioning. We use the total variation to rigorously formulate the continuum partitioning problem (1.4). The precise definition of the $\text{Cut}_\rho(A, A^c)$ in functional in (1.4) is

$$\text{Cut}_\rho(A, A^c) = TV(\mathbf{1}_A; D),$$

where $TV(\mathbf{1}_A; D)$ is defined in (3.1). We note that $TV(\mathbf{1}_A; D)$ is equal to $TV(\mathbf{1}_{A^c}; D)$, and is the perimeter of the set A in D weighted by ρ^2 .

We also formulate the balance terms, defined by (1.6) and (1.7), using characteristic functions. In fact, we start by extending the balance term to arbitrary functions $u \in L^1(\nu)$:

$$(3.7) \quad B_R(u) = \int_D |u(x) - \text{mean}_\rho(u)| \rho(x) \, dx \quad \text{and} \quad B_C(u) = \min_{c \in \mathbb{R}} \int_D |u(x) - c| \rho(x) \, dx,$$

where $\text{mean}_\rho(u)$ denotes the mean/expectation of $u(x)$ with respect to the measure $d\nu = \rho dx$. From here on, we use B to represent either B_R or B_C depending on the context. We have the relations:

$$(3.8) \quad B_R(\mathbf{1}_A) = \text{Bal}_R(A, A^c), \quad B_C(\mathbf{1}_A) = \text{Bal}_C(A, A^c),$$

for every measurable subset A of D . We also consider *normalized indicator functions* $\tilde{\mathbf{1}}_A$ given by

$$\tilde{\mathbf{1}}_A := \frac{\mathbf{1}_A}{B(\mathbf{1}_A)}, \quad A \subseteq D,$$

and consider the set

$$(3.9) \quad \text{Ind}(D) := \{u \in L^1(\nu) : u = \tilde{\mathbf{1}}_A \text{ for some measurable set } A \subseteq D \text{ with } B(\mathbf{1}_A) \neq 0\}.$$

Then for $u = \tilde{\mathbf{1}}_A \in \text{Ind}(D)$

$$(3.10) \quad TV(u) = TV(\tilde{\mathbf{1}}_A) = TV\left(\frac{\mathbf{1}_A}{B(\mathbf{1}_A)}\right) = \frac{TV(\mathbf{1}_A)}{B(\mathbf{1}_A)} = \frac{2\text{Cut}(A, A^c)}{\text{Bal}(A, A^c)}.$$

Thus, we deduce that problem (1.4) is equivalent to :

$$(3.11) \quad \text{Minimize} \quad E(u) := \begin{cases} TV(u) & \text{if } u \in \text{Ind}(D) \\ +\infty & \text{otherwise.} \end{cases}$$

Before we show that both the continuum ratio cut and Cheeger cut indeed have a minimizer we need the following lemma:

Lemma 6. (i) *The balance functions B are continuous on $L^1(\nu)$.*

(ii) *The set $\text{Ind}(D)$ is closed in $L^1(D)$.*

Proof. Let us start by proving (i). We first consider the balance term $B_C(u)$ that corresponds to the Cheeger Cut. Suppose that $u_k \rightarrow u$ in $L^1(\nu)$, and let c_k, c_∞ denote medians of u_k and u respectively. By definition, c_k and c satisfy

$$c_k \in \underset{c \in \mathbb{R}}{\text{argmin}} \int_D |u_k(x) - c| \rho(x) \, dx, \quad c_\infty \in \underset{c \in \mathbb{R}}{\text{argmin}} \int_D |u(x) - c| \rho(x) \, dx$$

This implies that

$$\int |u_k(x) - c_k| \rho(x) \, dx \leq \int |u_k(x) - c| \rho(x) \, dx$$

for any $c \in \mathbb{R}$, so that in particular we have

$$\begin{aligned} & \int |u_k - c_k| \rho(x) \, dx - \int |u - c_\infty| \rho(x) \, dx \\ & \leq \int |u_k - c_\infty| \rho(x) \, dx - \int |u - c_\infty| \rho(x) \, dx \leq \int |u_k - u| \rho(x) \, dx = \|u_k - u\|_{L^1(\nu)} \end{aligned}$$

Exchanging the role of u_k and u in this argument implies that the inequality

$$\int |u - c_\infty| \rho(x) \, dx - \int |u_k - c_k| \rho(x) \, dx \leq \int |u - u_k| \rho(x) \, dx \leq \|u_k - u\|_{L^1(\nu)}$$

also holds. Combining these inequalities shows that $|B(u_k) - B(u)| \leq \|u_k - u\|_{L^1(\nu)} \rightarrow 0$ as desired. Now consider the balance term $B_R(u)$ that corresponds to the ratio Cut. For the ratio cut, the inequality $||a| - |b|| \leq |a - b|$ immediately implies

$$\begin{aligned} & \left| \int |u_k - \text{mean}_\rho(u_k)| \rho(x) \, dx - \int |u - \text{mean}_\rho(u)| \rho(x) \, dx \right| \\ & \leq \int |u_k - u| \rho(x) \, dx + \int |\text{mean}_\rho(u_k) - \text{mean}_\rho(u)| \rho(x) \, dx \\ & \leq \int |u_k - u| \rho(x) \, dx + |\text{mean}_\rho(u_k) - \text{mean}_\rho(u)|. \end{aligned}$$

Since $u_k \rightarrow u$ in $L^1(\nu)$ we have that $\text{mean}_\rho(u_k) \rightarrow \text{mean}_\rho(u)$ and therefore $|B(u_k) - B(u)| \leq \|u_k - u\|_{L^1(\nu)} + |\text{mean}_\rho(u_k) - \text{mean}_\rho(u)| \rightarrow 0$ as desired.

In order to prove (ii) suppose that $\{u_k\}_{k \in \mathbb{N}}$ is a sequence in $\text{Ind}(D)$ converging in $L^1(\nu)$ to some $u \in L^1(\nu)$, we need to show that $u \in \text{Ind}(D)$. By (i) we know that $B(u_k) \rightarrow B(u)$ as $k \rightarrow \infty$. Since $u_k \in \text{Ind}(D)$, in particular $B(u_k) = 1$. Thus, $B(u) = 1$. On the other hand, $u_k \in \text{Ind}(D)$ implies that u_k has the form $u_k = \alpha_k \mathbf{1}_{A_k}$. Since this is true for every k , in particular we must have that u has the form $u = \alpha \mathbf{1}_A$ for some real number α and some measurable subset A of D . Finally, the fact that B is 1-homogeneous implies that $1 = B(u) = \alpha B(\mathbf{1}_A)$. In particular $B(\mathbf{1}_A) \neq 0$ and $\alpha = \frac{1}{B(\mathbf{1}_A)}$. Thus $u = \tilde{\mathbf{1}}_A$ with $B(\mathbf{1}_A) \neq 0$ and hence $u \in \text{Ind}(D)$. \square

Lemma 7. *Let D and ν be as stated at the beginning of this section. There exists a measurable set $A \subseteq D$ with $0 < \nu(A) < 1$ such that $\tilde{\mathbf{1}}_A$ minimizes (3.11).*

Proof. The statement follows by the direct method of the calculus of variations. Since the functional is bounded from below it suffices to show that it is lower semicontinuous with respect to the $L^1(\nu)$ norm and that a minimizing sequence is precompact in $L^1(\nu)$. To show lower semi-continuity it is enough to consider a sequence $u_n = \mathbf{1}_{A_n} \in \text{Ind}(D)$ converging in $L^1(\nu)$ to $u \in L^1(\nu)$. From Lemma 6 it follows that $u \in \text{Ind}(D)$ and hence $u = \tilde{\mathbf{1}}_A$ for some A with $B(A) > 0$. Therefore $\mathbf{1}_{A_n} \rightarrow \mathbf{1}_A$ as $n \rightarrow \infty$ in $L^1(\nu)$. The lower semi-continuity then follows from the lower semi-continuity of the total variation (3.6), the continuity of B and the fact that since $B(\mathbf{1}_A) > 0$, $1/B(\mathbf{1}_{A_n}) \rightarrow 1/B(\mathbf{1}_A)$ as $n \rightarrow \infty$.

The pre-compactness of any minimizing sequence of (3.11) follows directly from Theorem 5.1 in [6], which completes the proof. \square

4. ASSUMPTIONS AND STATEMENTS OF MAIN RESULTS.

Here we present the precise hypotheses we use and state precisely the main results of this paper. Let D be an open, bounded, connected subset of \mathbb{R}^d with Lipschitz boundary, and let $\rho : D \rightarrow \mathbb{R}$ be a continuous density which is bounded below and above by positive constants, that is, for all $x \in D$

$$(4.1) \quad \lambda \leq \rho(x) \leq \Lambda$$

for some $\Lambda \geq \lambda > 0$. We let ν be the measure $d\nu = \rho dx$. Let $\eta : [0, \infty) \rightarrow [0, \infty)$ be a similarity kernel, that is, a function satisfying:

(K1) $\eta(0) > 0$ and η is continuous at 0.

(K2) η is non-increasing.

(K3) $\sigma_\eta := \int_{\mathbb{R}^d} \eta(|x|) |\langle x, e_1 \rangle| \, dx < \infty$.

We refer to the quantity σ_η as the *surface tension* associated to η . These hypotheses on η hold for the standard similarity functions used in clustering contexts, such as the Gaussian similarity function $\eta(r) = \exp(-r^2)$ and the proximity similarity kernel (i.e. $\eta(r) = 1$ if $r \leq 1$ and $\eta(r) = 0$ otherwise).

The main result of our paper is:

Theorem 8 (Consistency of cuts). *Let domain D , probability measure ν and kernel η satisfy the conditions above. Let ε_n denote any sequence of positive numbers converging to zero that satisfy*

$$\lim_{n \rightarrow 0} \frac{(\log n)^{3/4}}{n^{1/2}} \frac{1}{\varepsilon_n} = 0 \quad (d = 2), \quad \lim_{n \rightarrow 0} \frac{(\log n)^{1/d}}{n^{1/d}} \frac{1}{\varepsilon_n} = 0 \quad (d \geq 3).$$

Let $\{\mathbf{x}_j\}_{j \in \mathbb{N}}$ be an i.i.d. sequence of points in D drawn from the density $\rho(x)$ and let $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$. Let $\mathcal{G}_n = (X_n, W_n)$ denote the graph whose edge weights are

$$w_{ij}^n := \eta \left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{\varepsilon_n} \right) \quad 1 \leq i, j \leq n.$$

Finally, let $\{Y_n^, Y_n^{*c}\}$ denote any optimal balanced cut of \mathcal{G}_n (solution of problem (1.1)). If $\{A^*, A^{*c}\}$ is the unique optimal balanced cut of the domain D (solution of problem (3.11)) then with probability one the sequence $\{Y_n^*, Y_n^{*c}\}$ converges to $\{A^*, A^{*c}\}$ in the TL^1 -sense. If there is more than one optimal continuum balanced cut (3.11) then $\{Y_n^*, Y_n^{*c}\}$ converges along a subsequence to an optimal continuum balanced cut.*

Additionally, \mathcal{C}_n , the minimum balance cut of the graph \mathcal{G}_n (the minimum of (1.1)), satisfies

$$(4.2) \quad \lim_{n \rightarrow \infty} \frac{\mathcal{C}_n}{n^2 \varepsilon_n^{d+1}} = \sigma_\eta \mathcal{C},$$

where σ_η is the surface tension associated to the kernel η and \mathcal{C} is the minimum of (1.4).

As we discussed in Remark 1 for $d \geq 3$ the scaling of ε on n is essentially the best possible.

The proof of Theorem 8 relies on establishing a variational convergence of discrete balanced cuts to continuum balanced cuts called the Γ -convergence which we recall in Subsection 5. The proof utilizes the results obtained in [18], where the notion of Γ -convergence is introduced in the context of data analysis problems, and in particular the Γ -convergence of the graph total variation is considered. The Γ -convergence, together with a compactness result, provides sufficient conditions for the convergence of minimizers of a given family of functionals to the minimizers of a limiting functional.

Remark 9. A few remarks help clarify the hypotheses and conclusions of our main result. The scaling condition $\varepsilon_n \gg (\log n)^{p_d} n^{-1/d}$ comes directly from the existence of transportation maps from Proposition (4). This means that ε_n must decay more slowly than the maximal distance a point in D has to travel to match its corresponding data point in X_n . In other words, the similarity graph \mathcal{G}_n must contain information on a larger scale than that on which the intrinsic randomness operates. Lastly, the conclusion of the theorem still holds if the partitions $\{Y_n^*, Y_n^{*c}\}$ only approximate an optimal balanced cut, that is if the energies of $\{Y_n^*, Y_n^{*c}\}$ satisfy

$$\lim_{n \rightarrow \infty} \left(\frac{\text{Cut}(Y_n^*, Y_n^{*c})}{\text{Bal}(Y_n^*, Y_n^{*c})} - \min_{Y \subseteq X_n} \frac{\text{Cut}(Y, Y^c)}{\text{Bal}(Y, Y^c)} \right) = 0.$$

This important property follows from a general result on Γ -convergence which we recall in Proposition 15.

We also establish the following multiclass equivalent to Theorem 8.

Theorem 10. *Let domain D , measure ν , kernel η , sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$, sample points $\{\mathbf{x}_i\}_{i \in \mathbb{N}}$, and graph \mathcal{G}_n satisfy the assumptions of Theorem 8. Let $(Y_1^{*n}, \dots, Y_R^{*n})$ denote any optimal balanced cut of \mathcal{G}_n , that is a minimizer of (1.3). If (A_1^*, \dots, A_R^*) is the unique optimal balanced cut of D (i.e.*

minimizer of (1.8)) then with probability one the sequence $(Y_1^{*n}, \dots, Y_R^{*n})$ converges to (A_1^*, \dots, A_R^*) in the TL^1 -sense. If the optimal continuum balanced cut is not unique then the convergence to a minimizer holds along subsequences. Additionally, \mathcal{C}_n , the minimum of (1.3), satisfies

$$\lim_{n \rightarrow \infty} \frac{\mathcal{C}_n}{n^2 \varepsilon_n^{d+1}} = \sigma_\eta \mathcal{C},$$

where σ_η is the surface tension associated to the kernel η and \mathcal{C} is the minimum of (1.8).

The proof of Theorem 10 involves modifying the geometric measure theoretical results from [18]. This leads to a substantially longer and more technical proof than the proof of Theorem 8, but the overall spirit of the proof remains the same in the sense that the Γ -convergence plays the leading role. Finally, we remark that analogous observations to the ones presented in Remark 9 apply to Theorem 10.

5. BACKGROUND ON Γ -CONVERGENCE

We recall and discuss the notion of Γ -convergence. In the literature Γ -convergence is defined for deterministic functionals. Nevertheless, the objects we are interested in are random and thus we decided to introduce this notion of convergence in this non-deterministic setting.

Let (X, d_X) be a metric space and let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space. Let $F_n : X \times \Omega \rightarrow [0, \infty]$ be a sequence of random functionals.

Definition 11. *The sequence $\{F_n\}_{n \in \mathbb{N}}$ Γ -converges with respect to metric d_X to the deterministic functional $F : X \rightarrow [0, \infty]$ as $n \rightarrow \infty$ if with \mathbb{P} -probability one the following conditions hold simultaneously:*

(1) **Liminf inequality:** *For every $x \in X$ and every sequence $\{x_n\}_{n \in \mathbb{N}}$ converging to x ,*

$$\liminf_{n \rightarrow \infty} F_n(x_n) \geq F(x),$$

(2) **Limsup inequality:** *For every $x \in X$ there exists a sequence $\{x_n\}_{n \in \mathbb{N}}$ converging to x satisfying*

$$\limsup_{n \rightarrow \infty} F_n(x_n) \leq F(x).$$

We say that F is the Γ -limit of the sequence of functionals $\{F_n\}_{n \in \mathbb{N}}$ (with respect to the metric d_X).

Remark 12. In most situations one does not prove the limsup inequality for all $x \in X$ directly. Instead, one proves the inequality for all x in a dense subset X' of X where it is somewhat easier to prove, and then deduce from this that the inequality holds for all $x \in X$. To be more precise, suppose that the limsup inequality is true for every x in a subset X' of X and the set X' is such that for every $x \in X$ there exists a sequence $\{x_k\}_{k \in \mathbb{N}}$ in X' converging to x and such that $F(x_k) \rightarrow F(x)$ as $k \rightarrow \infty$, then the limsup inequality is true for every $x \in X$. It is enough to use a diagonal argument to deduce this claim. This property is not related to the randomness of the functionals in any way.

Definition 13. *We say that the sequence of nonnegative random functionals $\{F_n\}_{n \in \mathbb{N}}$ satisfies the compactness property if with \mathbb{P} -probability one, the following statement holds: any sequence $\{x_n\}_{n \in \mathbb{N}}$ bounded in X and for which*

$$\sup_{k \in \mathbb{N}} F_n(x_n) < +\infty,$$

is relatively compact in X .

Remark 14. The boundedness assumption of $\{x_n\}_{n \in \mathbb{N}}$ in the previous definition is a necessary condition for relative compactness and so it is not restrictive.

The notion of Γ -convergence is particularly useful when the functionals $\{F_n\}_{n \in \mathbb{N}}$ satisfy the compactness property. This is because it guarantees that with \mathbb{P} -probability one, minimizers (or approximate minimizers) of F_n converge to minimizers of F and it also guarantees convergence of the minimum energy of F_n to the minimum energy of F (this statement is made precise in the next proposition). This is the reason why Γ -convergence is said to be a variational type of convergence. The next proposition can be found in [10, 16]. We present its proof for completeness and for the benefit of the reader. We also want to highlight the way this type of convergence works as ultimately this is one of the essential tools used to prove the main theorems of this paper.

Proposition 15. *Let $F_n : X \times \Omega \rightarrow [0, \infty]$ be a sequence of random nonnegative functionals which are not identically equal to $+\infty$, satisfying the compactness property and Γ -converging to the deterministic functional $F : X \rightarrow [0, \infty]$ which is not identically equal to $+\infty$. If it is true that with \mathbb{P} -probability one, there is a bounded sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfying*

$$(5.1) \quad \lim_{n \rightarrow \infty} \left(F_n(x_n) - \inf_{x \in X} F_n(x) \right) = 0$$

Then, with \mathbb{P} -probability one the following statements hold

$$(5.2) \quad \lim_{n \rightarrow \infty} \inf_{x \in X} F_n(x) = \min_{x \in X} F(x),$$

furthermore, every bounded sequence $\{x_n\}_{n \in \mathbb{N}}$ in X satisfying (5.1) is relatively compact and each of its cluster points is a minimizer of F . In particular, if F has a unique minimizer, then a bounded sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfying (5.1) converges to the unique minimizer of F .

Proof. Consider Ω' a set with \mathbb{P} -probability one for which all the statements in the definition of Γ -convergence together with the statement of the compactness property hold. We also assume that for every $\omega \in \Omega'$, there exists a bounded sequence $\{x_n\}_{n \in \mathbb{N}}$ satisfying (5.1). We fix such $\omega \in \Omega'$ and in particular we can assume that F_n is deterministic for every $n \in \mathbb{N}$.

Let $\{x_n\}_{n \in \mathbb{N}}$ be a sequence as the one described above. Let $\tilde{x} \in X$ be arbitrary. By the limsup inequality we know that there exists a sequence $\{\tilde{x}_n\}_{n \in \mathbb{N}}$ with $\tilde{x}_n \rightarrow \tilde{x}$ and such that

$$\limsup_{n \rightarrow \infty} F_n(\tilde{x}_n) \leq F(\tilde{x}).$$

By 5.1 we deduce that

$$(5.3) \quad \limsup_{n \rightarrow \infty} F(x_n) = \limsup_{n \rightarrow \infty} \inf_{x \in X} F_n(x) \leq \limsup_{n \rightarrow \infty} F_n(\tilde{x}_n) \leq F(\tilde{x}),$$

and since \tilde{x} was arbitrary we conclude that

$$(5.4) \quad \limsup_{n \rightarrow \infty} F_n(x_n) \leq \inf_{x \in X} F(x).$$

The fact that F is not identically equal to $+\infty$ implies that the term on the right hand side of the previous expression is finite and thus $\limsup_{n \rightarrow \infty} F_n(x_n) < +\infty$. Since the sequence $\{x_n\}_{n \in \mathbb{N}}$ was assumed bounded, we conclude from the compactness property for the sequence of functionals $\{F_n\}_{n \in \mathbb{N}}$ that $\{x_n\}_{n \in \mathbb{N}}$ is relatively compact.

Now let x^* be any accumulation point of the sequence $\{x_n\}_{n \in \mathbb{N}}$ (we know there exists at least one due to compactness), we want to show that x^* is a minimizer of F . Working along subsequences, we can assume without the loss of generality that $x_n \rightarrow x^*$. By the liminf inequality, we deduce that

$$(5.5) \quad \inf_{x \in X} F(x) \leq F(x^*) \leq \liminf_{n \rightarrow \infty} F(x_n).$$

The previous inequality and (5.3) imply that

$$F(x^*) \leq F(\tilde{x}),$$

where \tilde{x} is arbitrary. Thus, x^* is a minimizer of F and in particular $\inf_{x \in X} F(x) = \min_{x \in X} F(x)$. Finally, to establish (5.2) note that this follows from (5.4) and (5.5). \square

5.1. Γ -convergence of graph total variation. Of fundamental importance in obtaining our results is the Γ -convergence of the *graph total variation* proved in [18]. Let us describe this functional and also let us state the results we use. Given a point cloud $X_n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subseteq D$ where D is a domain in \mathbb{R}^d , we denote by $GTV_{n,\varepsilon_n} : L^1(\nu_n) \rightarrow [0, \infty]$ the functional:

$$(5.6) \quad GTV_{n,\varepsilon_n}(u_n) := \frac{1}{n^2 \varepsilon_n^{d+1}} \sum_{i,j=1}^n \eta \left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{\varepsilon_n} \right) |u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)|,$$

where η is a Kernel satisfying conditions **(K1)**–**(K3)**. The connection of the functional GTV_{n,ε_n} to problem (1.1) is the following: if Y_n is a subset of X_n , then the graph total variation of the indicator function $\mathbf{1}_{Y_n}$ is equal to a rescaled version of the graph cut of Y_n , that is,

$$GTV_{n,\varepsilon_n}(\mathbf{1}_{Y_n}) = \frac{2\text{Cut}(Y_n, Y_n^c)}{n^2 \varepsilon_n^{d+1}}.$$

Now we present the results obtained in [18].

Theorem 16 (Theorem 1.1 in [18]). *Let domain D , measure ν , kernel η , sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$, sample points $\{\mathbf{x}_i\}_{i \in N}$, and graph \mathcal{G}_n satisfy the assumptions of Theorem 8. Then, GTV_{n,ε_n} , defined by (5.6), Γ -converge to $\sigma_\eta TV$ as $n \rightarrow \infty$ in the TL^1 sense, where σ_η is the surface tension associated to the kernel η (see condition **(K3)**) and TV is the weighted (by ρ^2) total variation functional defined in (3.1).*

Moreover, we have the following compactness result.

Theorem 17 (Theorem 1.2 in [18]). *Under the same hypothesis of Theorem 1.1 in [18], the sequence of functionals $\{GTV_{n,\varepsilon_n}\}_{n \in \mathbb{N}}$ satisfies the compactness property. Namely, if a sequence $\{u_n\}_{n \in \mathbb{N}}$ with $u_n \in L^1(\nu_n)$ satisfies*

$$\sup_{n \in \mathbb{N}} \|u_n\|_{L^1(\nu_n)} < \infty,$$

and

$$\sup_{n \in \mathbb{N}} GTV_{n,\varepsilon_n}(u_n) < \infty,$$

then $\{u_n\}_{n \in \mathbb{N}}$ is TL^1 -relatively compact.

Finally, Corollary 1.3 in [18] allows us to restrict the functionals GTV_{n,ε_n} and TV to characteristic functions of sets and still obtain Γ -convergence.

Theorem 18 (Corollary 1.3 in [18]). *Under the assumptions of Theorem 1.1 in [18], with probability one the following statement holds: for every $A \subseteq D$ measurable, there exists a sequence of sets $\{Y_n\}_{n \in \mathbb{N}}$ with $Y_n \subseteq X_n$ such that,*

$$\mathbf{1}_{Y_n} \xrightarrow{TL^1} \mathbf{1}_A$$

and

$$\limsup_{n \rightarrow \infty} GTV_{n,\varepsilon_n}(\mathbf{1}_{Y_n}) \leq \sigma_\eta TV(\mathbf{1}_A).$$

6. CONSISTENCY OF TWO-WAY BALANCED CUTS

Here we prove Theorem 8.

6.1. Outline of the proof. Before proving that $\{Y_n^*, Y_n^{*c}\}$ converges to $\{A^*, A^{*c}\}$ in the sense of Definition 5, we first pause to outline the main ideas. Rather than directly working with the sets Y_n^* and Y_n^{*c} , it proved easier to work with their indicator functions $\mathbf{1}_{Y_n^*}(x)$ and $\mathbf{1}_{Y_n^{*c}}(x)$ instead. We first show, by an explicit construction in Subsection 6.2, that

$$(6.1) \quad u_n^* := \tilde{\mathbf{1}}_{Y_n^*}(x), \quad u_n^{**} := \tilde{\mathbf{1}}_{Y_n^{*c}}(x) \quad \text{minimize} \quad E_n(u_n) \quad \text{over all} \quad u_n \in L^1(\nu_n),$$

where E_n denotes a suitable objective function defined on $L^1(\nu_n)$, the set of functions defined over X_n . Each function $\tilde{\mathbf{1}}_{Y_n^*}(x) := \alpha_n \mathbf{1}_{Y_n^*}(x)$ is simply a rescaled version of the original indicator function (for some explicit coefficient α_n that we will define later). Similarly, in Subsection 3.2 we showed that the normalized indicator functions

$$(6.2) \quad u^* := \tilde{\mathbf{1}}_{A^*}(x), \quad u^{**} := \tilde{\mathbf{1}}_{A^{*c}}(x) \quad \text{minimize} \quad E(u) \quad \text{over all} \quad u \in L^1(\nu),$$

where E is defined by (3.11).

In Subsection 6.3 we show that the approximating functionals E_n Γ -converge to $\sigma_\eta E$ in the TL^1 -sense. In Lemma 21 we establish that u_n^* and u_n^{**} exhibit the required compactness. Thus, they must converge toward the normalized indicator functions $\tilde{\mathbf{1}}_{A^*}(x)$ and $\tilde{\mathbf{1}}_{A^{*c}}(x)$ up to relabeling (see Proposition 15). If $\{A^*, A^{*c}\}$ is the unique minimizer, the convergence of the whole sequence follows. The convergence of the partition $\{Y_n^*, Y_n^{*c}\}$ toward the partition $\{A^*, A^{*c}\}$ in the sense of Definition 5 is a direct consequence. The convergence (4.2) follows from (5.2) in Proposition 15.

6.2. Functional description of discrete cuts. We introduce functionals that describe the discrete ratio and Cheeger cuts in terms of functions on X_n , rather than in terms of subsets of X_n . This mirrors the description of continuum partitions provided in Subsection 3.2. For $u_n \in L^1(\nu_n)$, we start by defining

$$(6.3) \quad B_R^n(u_n) := \frac{1}{n} \sum_{i=1}^n |u_n(\mathbf{x}_i) - \text{mean}_n(u_n)| \quad \text{and} \quad B_C^n(u_n) := \min_{c \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |u_n(\mathbf{x}_i) - c|.$$

Here $\text{mean}_n(u_n) = \frac{1}{n} \sum_{i=1}^n u_n(\mathbf{x}_i)$. A straightforward computation shows that for $Y_n \subseteq X_n$

$$(6.4) \quad B_R^n(\mathbf{1}_{Y_n}) = \text{Bal}_R(Y_n, Y_n^c), \quad B_C^n(\mathbf{1}_{Y_n}) = \text{Bal}_C(Y_n, Y_n^c).$$

From here on we write B_n to represent either B_R^n or B_C^n depending on the context.

Instead of defining $E_n(u_n)$ simply as the ratio $GTV_{n,\varepsilon_n}(u_n)/B_n(u_n)$, which is the direct analogue of (1.1), it proves easier to work with suitably normalized indicator functions. Given $Y_n \subseteq X_n$ with $B_n(\mathbf{1}_{Y_n}) \neq 0$, the *normalized indicator function* $\tilde{\mathbf{1}}_{Y_n}(x)$ is defined by

$$\tilde{\mathbf{1}}_{Y_n}(x) = \mathbf{1}_{Y_n}(x)/B_C^n(\mathbf{1}_{Y_n}) \quad \text{or} \quad \tilde{\mathbf{1}}_{Y_n}(x) = \mathbf{1}_{Y_n}(x)/B_R^n(\mathbf{1}_{Y_n}).$$

Note that $B_n(\tilde{\mathbf{1}}_A) = 1$. We also restrict the minimization of $E_n(u)$ to the set

$$(6.5) \quad \text{Ind}_n(D) := \{u_n \in L^1(\nu_n) : u_n = \tilde{\mathbf{1}}_{Y_n} \text{ for some } Y_n \subseteq X_n \text{ with } B_n(\mathbf{1}_{Y_n}) \neq 0\}.$$

Now, suppose that $u_n \in \text{Ind}_n(D)$, i.e. that $u_n = \tilde{\mathbf{1}}_{Y_n}$ for some set Y_n with $B_n(\mathbf{1}_{Y_n}) > 0$. Using (3.8) together with the fact that GTV_{n,ε_n} (defined in (5.6)) is one-homogeneous implies, as in (3.10)

$$(6.6) \quad GTV_{n,\varepsilon_n}(u_n) = \frac{2}{n^2 \varepsilon_n^{d+1}} \frac{\text{Cut}(Y_n, Y_n^c)}{\text{Bal}(Y_n, Y_n^c)}.$$

Thus, minimizing GTV_{n,ε_n} over all $u_n \in \text{Ind}_n(D)$ is equivalent to the balanced graph-cut problem (1.1) on the graph $\mathcal{G}_n = (X_n, W_n)$ constructed from the first n data points. We have therefore arrived at our destination, i.e. a proper reformulation of (1.1) defined over functions $u_n \in L^1(\nu_n)$ instead of subsets of X_n :

$$(6.7) \quad \text{Minimize} \quad E_n(u_n) := \begin{cases} GTV_{n,\varepsilon_n}(u_n) & \text{if } u_n \in \text{Ind}_n(D) \\ +\infty & \text{otherwise.} \end{cases}$$

6.3. Γ -Convergence.

Proposition 19. (Γ -Convergence) *Let domain D , measure ν , kernel η , sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$, sample points $\{\mathbf{x}_i\}_{i \in N}$, and graph \mathcal{G}_n satisfy the assumptions of Theorem 8. Let E_n be as defined in (6.7) and E as in (3.11). Then*

$$E_n \xrightarrow{\Gamma} \sigma_\eta E \quad \text{with respect to } TL^1 \text{ metric as } n \rightarrow \infty$$

where σ_η is the surface tension defined in assumption (K3). That is

$$(1) \text{ For any } u \in L^1(\nu) \text{ and any sequence } \{u_n\}_{n \in \mathbb{N}} \text{ with } u_n \in L^1(\nu_n) \text{ that converges to } u \text{ in } TL^1, \\ (6.8) \quad \sigma_\eta E(u) \leq \liminf_{n \rightarrow \infty} E_n(u_n).$$

$$(2) \text{ For any } u \in L^1(\nu) \text{ there exists at least one sequence } \{u_n\}_{n \in \mathbb{N}} \text{ with } u_n \in L^1(\nu_n) \text{ which} \\ \text{converges to } u \text{ in } TL^1 \text{ and also satisfies} \\ (6.9) \quad \limsup_{n \rightarrow \infty} E_n(u_n) \leq \sigma_\eta E(u).$$

We leverage Theorem 16 to prove this claim. We first need a preliminary lemma which allows us to handle the presence of the additional balance terms in (6.7) and (3.11).

Lemma 20. (i) *If $\{u_n\}_{n \in \mathbb{N}}$ is a sequence with $u_n \in L^1(\nu_n)$ and $u_n \xrightarrow{TL^1} u$ for some $u \in L^1(\nu)$, then $B_n(u_n) \rightarrow B(u)$.*
 (ii) *If $u_n = \tilde{\mathbf{1}}_{Y_n}$, where $Y_n \subset X_n$, converges to $u = \tilde{\mathbf{1}}_A$ in the TL^1 -sense, then $\mathbf{1}_{Y_n}$ converges to $\mathbf{1}_A$ in the TL^1 -sense.*

Proof. To prove (i), suppose that $u_n \in L^1(\nu_n)$ and that $u_n \xrightarrow{TL^1} u$. Let us consider $\{T_n\}_{n \in \mathbb{N}}$ a stagnating sequence of transportation maps between ν and $\{\nu_n\}_{n \in \mathbb{N}}$. Then, we have $u_n \circ T_n \xrightarrow{L^1(\nu)} u$ and thus by (i), we have that $B(u_n \circ T_n) \rightarrow B(u)$. To conclude the proof we notice that $B(u_n \circ T_n) = B_n(u_n)$ for every n . In fact, by the change of variables (2.2) we have that for every $c \in \mathbb{R}$

$$(6.10) \quad \int_D |u_n(x) - c| d\nu_n(x) = \int_D |u_n \circ T_n(x) - c| d\nu(x)$$

In particular we have $B_C^n(u_n) = B_C(u_n \circ T_n)$. Applying the change of variables (2.2), we obtain $\text{mean}_n(u_n) = \text{mean}_\rho(u_n \circ T_n)$ and combining with (6.10) we deduce that $B_R^n(u_n) = B_R(u_n \circ T_n)$.

The proof of (ii) is straightforward. \square

Now we turn to the proof of Proposition 19.

Proof. **Liminf Inequality.** For arbitrary $u \in L^1(\nu)$ and arbitrary sequence $\{u_n\}_{n \in \mathbb{N}}$ with $u_n \in L^1(\nu_n)$ and with $u_n \xrightarrow{TL^1} u$, we need to show that

$$\liminf_{n \rightarrow \infty} E_n(u_n) \geq \sigma_\eta E(u).$$

First assume that $u \in \text{Ind}(D)$. In particular $E(u) = TV(u)$. Now, note that working along a subsequence we can assume that the liminf is actually a limit and that this limit is finite (otherwise the inequality would be trivially satisfied). This implies that for all n large enough we have $E(u_n) < +\infty$, which in particular implies that $E_n(u_n) = GTV_{n, \varepsilon_n}(u_n)$. Theorem 16 then implies that

$$\liminf_{n \rightarrow \infty} E_n(u_n) = \liminf_{n \rightarrow \infty} GTV_{n, \varepsilon_n}(u_n) \geq \sigma_\eta TV(u) = \sigma_\eta E(u).$$

Now let us assume that $u \notin \text{Ind}(D)$. Let us consider a stagnating sequence of transportation maps $\{T_n\}_{n \in \mathbb{N}}$ between $\{\nu_n\}_{n \in \mathbb{N}}$ and ν . Since $u_n \xrightarrow{TL^1} u$ then $u_n \circ T_n \xrightarrow{L^1(\nu)} u$. By Lemma 20, the set

$\text{Ind}(D)$ is a closed subset of $L^1(\nu)$. We conclude that $u_n \circ T_n \notin \text{Ind}(D)$ for all large enough n . From the proof of Lemma 20 we know that $B_n(u_n) = B(u_n \circ T_n)$ and from this fact, it is straightforward to show that $u_n \circ T_n \notin \text{Ind}(D)$ if and only if $u_n \notin \text{Ind}_n(D)$. Hence, $u_n \notin \text{Ind}_n(D)$ for all large enough n and in particular $\liminf_{n \in \mathbb{N}} E_n(u_n) = +\infty$ which implies that the desired inequality holds in this case.

Limsup Inequality. We now consider $u \in L^1(\nu)$. We want to show that there exists a sequence $\{u_n\}_{n \in \mathbb{N}}$ with $u_n \in L^1(\nu_n)$ such that

$$\limsup_{n \rightarrow \infty} E_n(u_n) \leq \sigma_\eta E(u).$$

Let us start by assuming that $u \notin \text{Ind}(D)$. In this case $E(u) = +\infty$. From Theorem 16 we know there exists at least one sequence $\{u_n\}_{n \in \mathbb{N}}$ with $u_n \in L^1(\nu_n)$ such that $u_n \xrightarrow{TL^1} u$. Since $E(u) = +\infty$, the inequality is trivially satisfied in this case.

On the other hand, if $u \in \text{Ind}(D)$, we know that $u = \mathbf{1}_A$ for some measurable subset A of D with $B(\mathbf{1}_A) \neq 0$. By Theorem 18, there exists a sequence $\{Y_n\}_{n \in \mathbb{N}}$ with $Y_n \subseteq X_n$, satisfying $\mathbf{1}_{Y_n} \xrightarrow{TL^1} \mathbf{1}_A$ and

$$(6.11) \quad \limsup_{n \rightarrow \infty} GTV_{n, \varepsilon_n}(\mathbf{1}_{Y_n}) \leq \sigma_\eta TV(\mathbf{1}_A).$$

Since $\mathbf{1}_{Y_n} \xrightarrow{TL^1} \mathbf{1}_A$ Lemma 20 implies that

$$(6.12) \quad B_n(\mathbf{1}_{Y_n}) \rightarrow B(\mathbf{1}_A).$$

In particular $B_n(\mathbf{1}_{Y_n}) \neq 0$ for all n large enough, and thus we can consider the function $u_n := \tilde{\mathbf{1}}_{Y_n} \in \text{Ind}_n(D)$. From (6.12) it follows that $u_n \xrightarrow{TL^1} u$ and together with (6.11) it follows that

$$\limsup_{n \rightarrow \infty} GTV_{n, \varepsilon_n}(u_n) = \limsup_{n \rightarrow \infty} \frac{1}{B_n(Y_n)} GTV_{n, \varepsilon_n}(\mathbf{1}_{Y_n}) \leq \frac{1}{B(\mathbf{1}_A)} \sigma_\eta TV(\mathbf{1}_A) = \sigma_\eta TV(u)$$

Since, $u_n \in \text{Ind}_n(D)$ for all n large enough, in particular we have $GTV_{n, \varepsilon_n}(\mathbf{1}_{Y_n}) = E_n(\mathbf{1}_{Y_n})$ and also since $u \in \text{Ind}(D)$, we have $E(u) = TV(u)$. These facts together with the previous chain of inequalities imply the result. \square

6.4. Compactness.

Lemma 21 (Compactness). *Any subsequence of $\{u_n^*\}_{n \geq 1}$ or $\{u_n^{**}\}_{n \geq 1}$ of minimizers of E_n (defined in (6.1) and (6.2)) has a further subsequence that converges in the TL^1 -sense.*

Proof. Let u_n^*, u_n^{**} denote minimizing sequences. To show that any subsequence of u_n^* has a convergent subsequence, it suffices to show that both

$$(6.13) \quad \limsup_{n \rightarrow \infty} GTV_{n, \varepsilon_n}(u_n^*) < +\infty$$

$$(6.14) \quad \limsup_{n \rightarrow \infty} \|u_n^*\|_{L^1(\nu_n)} < +\infty$$

hold due to Theorem 1.2 in [18]. From the Γ -convergence established in Proposition 19 and from the proof of Proposition 15 it follows that (6.13) is satisfied for both minimizing sequences. Recall that $u_n^* = \mathbf{1}_{Y_n^*}/B_n(\mathbf{1}_{Y_n^*})$ and that $u_n^{**} = \mathbf{1}_{Y_n^{*c}}/B_n(\mathbf{1}_{Y_n^{*c}})$, where Y_n^* denotes an optimal balanced cut.

To show (6.14), consider first the balance term that corresponds to the Cheeger Cut. Define a sequence v_n as follows. Set $v_n := u_n^*$ if $|Y_n^*| \leq |Y_n^{*c}|$ and $v_n = u_n^{**}$ otherwise. It then follows that

$$\|v_n\|_{L^1(\nu_n)} = \frac{\min\{|Y_n^*|, |Y_n^{*c}|\}}{\min\{|Y_n^*|, |Y_n^{*c}|\}} = 1.$$

Also, note that $GTV_{n,\varepsilon_n}(v_n) = GTV_{n,\varepsilon_n}(u_n^*)$. Thus (6.13) and (6.14) hold for v_n , so that any subsequence of v_n has a convergent subsequence in the TL^1 -sense. Let $v_{n_k} \xrightarrow{TL^1} v$ denote a convergent subsequence. Now observe that by construction v_{n_k} minimizes E_{n_k} for every k . Thus, it follows from Proposition 19 and general properties of Γ -convergence (see Proposition 15), that v minimizes E and in particular v is a normalized characteristic function, that is, $v = \mathbf{1}_A/B(\mathbf{1}_A)$ for some $A \subseteq D$ with $B(\mathbf{1}_A) \neq 0$. Since $B_{n_k}(\mathbf{1}_{Y_{n_k}^*}) = B_{n_k}(\mathbf{1}_{Y_{n_k}^{*c}})$, $v_{n_k} \xrightarrow{TL^1} v$ implies that

$$\frac{1}{B_{n_k}(Y_{n_k}^*)} \rightarrow \frac{1}{B(A)}.$$

Therefore, for large enough k we have

$$\|u_{n_k}^*\|_{L^1(\nu_{n_k})} \leq \frac{1}{B_{n_k}(Y_{n_k}^*)} \leq \frac{2}{B(A)}$$

and

$$\|u_{n_k}^{**}\|_{L^1(\nu_{n_k})} \leq \frac{1}{B_{n_k}(Y_{n_k}^{*c})} = \frac{1}{B_{n_k}(Y_{n_k}^*)} \leq \frac{2}{B(A)}.$$

We conclude that $\|u_{n_k}^*\|_{L^1(\nu_{n_k})}$ and $\|u_{n_k}^{**}\|_{L^1(\nu_{n_k})}$ remain bounded, so that both minimizing subsequences satisfy (6.14) and (6.13) simultaneously. This yields compactness in the Cheeger Cut case.

Now consider the balance term $B(u) = B_R(u)$ that corresponds to the Ratio Cut. Define a sequence $v_n := u_n^* - \text{mean}_n(u_n^*)$, and note that $GTV_{n,\varepsilon_n}(v_n) = GTV_{n,\varepsilon_n}(u_n^*)$ since the total variation is invariant with respect to translation. It then follows that

$$\|v_n\|_{L^1(\nu)} = \int_D |u_n^*(x) - \text{mean}_\rho(u_n^*)| \rho(x) \, dx = B(u_n^*) = 1.$$

Thus the sequence $\{v_n\}_{n \in \mathbb{N}}$ is precompact in TL^1 . Let $v_{n_k} \xrightarrow{TL^1} v$ denote a convergent subsequence. Using a stagnating sequence of transportation maps $\{T_{n_k}\}_{k \in \mathbb{N}}$ between ν and the sequence of measures $\{\nu_{n_k}\}_{k \in \mathbb{N}}$, we have that $v_{n_k} \circ T_{n_k} \xrightarrow{L^1(\nu)} v$. By passing to a further subsequence if necessary, we may assume that $v_{n_k} \circ T_{n_k}(x) \rightarrow v(x)$ for almost every x in D .

For any such x , we have that either $T_{n_k}(x) \in Y_{n_k}^*$ or $T_{n_k}(x) \in Y_{n_k}^{*c}$ so that either

$$v_{n_k} \circ T_{n_k}(x) = \frac{1}{2|Y_{n_k}^*|} \quad \text{or} \quad v_{n_k} \circ T_{n_k}(x) = -\frac{1}{2|Y_{n_k}^{*c}|}.$$

Now, by continuity of the balance term, we have

$$B(v) = \lim_{k \rightarrow \infty} B_{n_k}(v_{n_k}) = 1,$$

and also

$$\text{mean}_\rho(v) = \lim_{k \rightarrow \infty} \text{mean}_{n_k}(v_{n_k}) = 0.$$

In particular the measure of the region in which v is positive is strictly greater than zero, and likewise the measure of the region in which v is negative is strictly greater than zero. It follows that both $|Y_{n_k}^*|$ and $|Y_{n_k}^{*c}|$ remain bounded away from zero for all k sufficiently large. As a consequence, the fact that

$$\|u_{n_k}^*\|_{L^1(\nu_{n_k})} = \frac{1}{2|Y_{n_k}^{*c}|}, \quad \|u_{n_k}^{**}\|_{L^1(\nu_{n_k})} = \frac{1}{2|Y_{n_k}^*|},$$

implies that both (6.13) and (6.14) hold along a subsequence, yielding the desired compactness. \square

6.5. Conclusion of the proof of Theorem 8. We may now turn to the final step of the proof. From Proposition 15, we know that any limit point of $\{u_n^*\}_{n \in \mathbb{N}}$ (in the TL^1 sense) must equal u^* or u^{**} . As a consequence, for any subsequence $u_{n_k}^*$ that converges to u^* we have that $\mathbf{1}_{Y_{n_k}^*} \xrightarrow{TL^1} \mathbf{1}_{A^*}$ by lemma 20, while $\mathbf{1}_{Y_{n_k}^{*c}} \xrightarrow{TL^1} \mathbf{1}_{A^{*c}}$ if the subsequence converges to u^{**} instead. Moreover, in the first case we would also have $\mathbf{1}_{Y_{n_k}^{*c}} \xrightarrow{TL^1} \mathbf{1}_{A^{*c}}$ and in the second case $\mathbf{1}_{Y_{n_k}^*} \xrightarrow{TL^1} \mathbf{1}_{A^*}$. Thus in either case we have

$$\{Y_{n_k}^*, Y_{n_k}^{*c}\} \xrightarrow{TL^1} \{A^*, A^{*c}\}$$

Thus, for any subsequence of $\{Y_n^*, Y_n^{*c}\}_{n \in \mathbb{N}}$ it is possible to obtain a further subsequence converging to $\{A^*, A^{*c}\}$, and thus the full sequence converges to $\{A^*, A^{*c}\}$.

7. CONSISTENCY OF MULTIWAY BALANCED CUTS

Here we prove Theorem 10.

Just as what we did in the two-class case, the first step in the proof of Theorem 10 involves a reformulation of both the balanced graph-cut problem (1.3) and the analogous balanced domain-cut problem (1.8) as equivalent minimizations defined over spaces of functions and not just spaces of partitions or sets.

We let $B_n(u_n) := \text{mean}_n(u_n)$ for $u_n \in L^1(\nu_n)$ and $B(u) := \text{mean}_\rho(u)$ for $u \in L^1(\nu)$, to be the corresponding balance terms. Given this balance terms, we let $\text{Ind}_n(D)$ and $\text{Ind}(D)$ be defined as in (6.5) and (3.9) respectively.

We can then let the sets $\mathcal{M}_n(D)$ and $\mathcal{M}(D)$ to consist of those collections $\mathcal{U} = (u_1, \dots, u_R)$ comprised of exactly R disjoint, normalized indicator functions that cover D . The sets $\mathcal{M}_n(D)$ and $\mathcal{M}(D)$ are the multi-class analogues of $\text{Ind}_n(D)$ and $\text{Ind}(D)$ respectively. Specifically, we let

(7.1)

$$\mathcal{M}_n(D) = \left\{ (u_1^n, \dots, u_R^n) : u_r^n \in \text{Ind}_n(D), \int_D u_r^n(x) u_s^n(x) d\nu_n(x) = 0 \text{ if } r \neq s, \sum_{r=1}^R u_r^n > 0 \right\}$$

(7.2)

$$\mathcal{M}(D) = \left\{ (u_1, \dots, u_R) : u_r \in \text{Ind}(D), \int_D u_r(x) u_s(x) d\nu(x) = 0 \text{ if } r \neq s, \sum_{r=1}^R u_r > 0 \right\}.$$

Note for example that if $\mathcal{U} = (u_1, \dots, u_R) \in \mathcal{M}(D)$, then the functions u_r are normalized indicator functions, $u_r = \mathbf{1}_{A_r}/|A_r|$ for $1 \leq r \leq R$, and the orthogonality constraints imply that $\{A_1, \dots, A_R\}$ is a collection of pairwise disjoint sets (up to Lebesgue-null sets). Additionally, the condition that $\sum_{r=1}^R u_r > 0$ holds almost everywhere implies that the sets $\{A_1, \dots, A_R\}$ cover D up to Lebesgue-null sets.

With these definitions in hand, we may follow the same argument in the two-class case to conclude that that the minimization

$$(7.3) \quad \text{Minimize} \quad E_n(\mathcal{U}_n) := \begin{cases} \sum_{r=1}^R GTV_{n, \varepsilon_n}(u_r^n) & \text{if } \mathcal{U}_n \in \mathcal{M}_n(D) \\ +\infty & \text{otherwise} \end{cases}$$

is equivalent to the balanced graph-cut problem (1.3), while the minimization

$$(7.4) \quad \text{Minimize} \quad E(\mathcal{U}) := \begin{cases} \sum_{r=1}^R TV(u_r) & \text{if } \mathcal{U} \in \mathcal{M}(D) \\ +\infty & \text{otherwise} \end{cases}$$

is equivalent to the balance domain-cut problem (1.8).

At this stage, the proof of Theorem 10, is completed by following the same steps as in the two-class case. In particular we want to show that E_n defined in (7.3) Γ -converges in the TL^1 -sense to $\sigma_\eta E$, where E is defined in 7.4. That is, we want to prove the following.

Proposition 22. (Γ -Convergence) *Let domain D , measure ν , kernel η , sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$, sample points $\{\mathbf{x}_i\}_{i \in N}$, and graph \mathcal{G}_n satisfy the assumptions of Theorem 8. Consider functionals E_n of (7.3) and E of (7.4). Then*

$$E_n \xrightarrow{\Gamma} \sigma_n E \quad \text{with respect to } (TL^1)^{\otimes R} \text{ metric as } n \rightarrow \infty.$$

That is

(1) *For any $\mathcal{U} \in [L^1(D)]^R$ and any sequence $\mathcal{U}_n \in (L^1(\nu_n))^R$ that converges to \mathcal{U} in the TL^1 -sense,*

$$(7.5) \quad E(\mathcal{U}) \leq \liminf_{n \rightarrow \infty} E_n(\mathcal{U}_n).$$

(2) *For any $\mathcal{U} \in [L^1(D)]^R$ there exists at least one sequence \mathcal{U}_n that both converges to \mathcal{U} in the TL^1 -sense and also satisfies*

$$(7.6) \quad \limsup_{n \rightarrow \infty} E_n(\mathcal{U}_n) \leq E(\mathcal{U}).$$

Remark 23. We remark that all the types of convergence for vector-valued functions are to be understood as component-wise convergence in the corresponding topology. This helps us clarify the way the TL^1 -convergence is considered in Proposition 22.

Assuming Proposition 22, the following lemma follows in a straightforward way. We omit its proof since it follows analogous arguments to the ones used in the proof of Proposition 21.

Lemma 24 (Compactness). *Any subsequence of $\{\mathcal{U}_n^*\}_{n \geq 1}$ of minimizers to (7.3) has a further subsequence that converges in the TL^1 -sense.*

Finally, due to Proposition 22 and Lemma 24, the arguments presented in Subsections 6.1 and 6.5 can be adapted in a straightforward way to complete the proof of Theorem 10. So we focus on the proof of Proposition 22, where arguments not present in the two-class case are needed. On one hand, this is due to the presence of the orthogonality constraints in the definition of $\mathcal{M}_n(D)$ and $\mathcal{M}(D)$, and on the other hand, from a geometric measure theory perspective, due to the fact that an arbitrary partition of the domain D into more than two sets can not be approximated by smooth partitions as multiple junctions appear when more than two sets in the partition meet.

7.1. Proof of Proposition 22. The next lemma is the multiclass analogue of Lemmas 6 and 20 combined.

Lemma 25. (i) *If $\mathcal{U}_k \rightarrow \mathcal{U}$ in $(L^1(\nu))^R$ then $B(u_r^k) \rightarrow B(u_r)$ for all $1 \leq r \leq R$.* (ii) *The set $\mathcal{M}(D)$ is closed in $L^1(\nu)$.* (iii) *If $\{\mathcal{U}_n\}$ is a sequence with $\mathcal{U}_n \in (L^1(\nu_n))^R$ and $\mathcal{U}_n \xrightarrow{TL^1} \mathcal{U}$ for some $\mathcal{U} \in (L^1(\nu))^R$, then $B_n(u_r^n) \rightarrow B(u_r)$ for all $1 \leq r \leq R$.* (iv) *If $u_n = \tilde{\mathbf{1}}_{Y_n}$, where $Y_n \subset X_n$, converges to $u = \tilde{\mathbf{1}}_A$ in the TL^1 -sense, then $\mathbf{1}_{Y_n}$ converges to $\mathbf{1}_A$ in the TL^1 -sense.*

Proof. Statements (i), (iii) and (iv) follow directly from the proof of Proposition 20.

In order to prove the second statement, suppose that a sequence $\{\mathcal{U}_k\}_{k \in \mathbb{N}}$ in $\mathcal{M}(D)$ converges to some \mathcal{U} in $(L^1(\nu))^R$. We need to show that $\mathcal{U} \in \mathcal{M}(D)$. First of all note that for every $1 \leq r \leq R$, $u_r^k \xrightarrow{L^1(\nu)} u_r$. Since $u_r^k \in \text{Ind}(D)$ for every $k \in \mathbb{N}$, and since $\text{Ind}(D)$ is a closed subset of $L^1(\nu)$ (by Proposition 20), we deduce that $u_r \in \text{Ind}(D)$ for every r .

The orthogonality condition follows from Fatou's lemma. In fact, working along a subsequence we can without the loss of generality assume that for every r , $u_r^k \rightarrow u_r$ for almost every x in D . Hence, for $r \neq s$ we have

$$0 \leq \int_D u_r(x) u_s(x) d\nu(x) = \int_D \liminf_{k \rightarrow \infty} (u_r^k(x) u_s^k(x)) d\nu(x) \leq \liminf_{k \rightarrow \infty} \int_D u_r^k(x) u_s^k(x) d\nu(x) = 0$$

Now let us write $u_r^k = \mathbf{1}_{A_r^k} / B(\mathbf{1}_{A_r^k})$ and $u_r = \mathbf{1}_{A_r} / B(\mathbf{1}_{A_r})$. As in the proof of Proposition (20) we must have $B(\mathbf{1}_{A_r^k}) \rightarrow B(\mathbf{1}_{A_r})$ as $k \rightarrow \infty$. Thus, for almost every $x \in D$

$$\sum_{r=1}^R u_r(x) = \lim_{k \rightarrow \infty} \sum_{r=1}^R u_r^k(x) \geq \lim_{k \rightarrow \infty} \min_{r=1, \dots, R} \frac{1}{B(\mathbf{1}_{A_r^k})} = \min_{r=1, \dots, R} \frac{1}{B(\mathbf{1}_{A_r})} > 0.$$

□

Proof of Proposition 22. Liminf inequality. The proof of (7.5) follows the approach used in the two-class case. Let $\mathcal{U}_n \xrightarrow{TL^1} \mathcal{U}$ denote an arbitrary convergent sequence. As $\mathcal{M}(D)$ is closed, if $\mathcal{U} \notin \mathcal{M}(D)$ then as in the two-class case, it is easy to see that $\mathcal{U}_n \notin \mathcal{M}_n(D)$ for all n sufficiently large. The inequality (7.5) is then trivial in this case, as both sides of it are equal to infinity. Conversely, if $\mathcal{U} \in \mathcal{M}(D)$ then we may assume that $\mathcal{U}_n \in \mathcal{M}_n(D)$ for all n , since only those terms with $\mathcal{U}_n \in \mathcal{M}_n(D)$ can contribute non-trivially to the limit inferior. In this case we easily have

$$\begin{aligned} \liminf_{n \rightarrow \infty} E_n(\mathcal{U}_n) &= \liminf_{n \rightarrow \infty} \sum_{r=1}^R GTV_{n, \varepsilon_n}(u_r^n) \geq \sum_{r=1}^R \liminf_{n \rightarrow \infty} GTV_{n, \varepsilon_n}(u_r^n) \\ &\geq \sigma_\eta \sum_{r=1}^R TV(u_r) = \sigma_\eta E(\mathcal{U}). \end{aligned}$$

The last inequality follows from Theorem 16. This establishes the first statement in Proposition 22.

Limsup inequality. We now turn to the proof of (7.6), which is significantly more involved than the two-class argument due to the presence of the orthogonality constraints. Borrowing terminology from the Γ -convergence literature, we say that $\mathcal{U} \in (L^1(\nu))^R$ has a *recovery sequence* when there exists a sequence $\mathcal{U}_n \in (L^1(\nu_n))^R$ such that (7.6) holds. To show that each $\mathcal{U} \in (L^1(\nu))^R$ has a recovery sequence, we first remark that due to general properties of the Γ -convergence, it is enough to verify (7.6) for \mathcal{U} belonging to a dense subset of $\mathcal{M}(D)$ with respect to the energy E (see Remark 12). We furthermore remark that it is enough to consider $\mathcal{U} = (u_1, \dots, u_R) \in (L^1(D))^R$ for which $E(\mathcal{U}) < \infty$, as the other case is trivial. So we can consider $\mathcal{U} \in \mathcal{M}(D)$ that satisfy

$$\sum_{r=1}^R TV(u_r) < \infty.$$

Let $u_r = \mathbf{1}_{A_r} / B(\mathbf{1}_{A_r})$ and let $c_0 := \max\{B(\mathbf{1}_{A_1}), \dots, B(\mathbf{1}_{A_R})\}$ denote the size of the largest set in the collection. The fact that $E(\mathcal{U}) < \infty$ then implies

$$TV(\mathbf{1}_{A_r}) \leq c_0 TV(u_r) \leq c_0 \sum_{r=1}^R TV(u_r) < \infty,$$

so that all sets $\{A_1, \dots, A_R\}$ in the collection defining \mathcal{U} have finite perimeter. Additionally because $\mathcal{U} \in \mathcal{M}(D)$ implies that any two sets A_r, A_s with $r \neq s$ have empty intersection up to a Lebesgue-null set, we may freely assume without the loss of generality that the sets $\{A_1, \dots, A_R\}$ are mutually disjoint.

Let us define sets with *piecewise (PW) smooth* boundary to be the subsets of \mathbb{R}^d whose boundary is a subset of finitely many open $d - 1$ -dimensional manifolds embedded in \mathbb{R}^d . We start by constructing a recovery sequence for a \mathcal{U} whose defining sets $\{A_1, \dots, A_R\}$ are of the form $A_r = B_r \cap D$, where B_r has piecewise smooth boundary and satisfies $|D\mathbf{1}_{B_r}|_{\rho^2}(D) = 0$. We say that such \mathcal{U} is *induced by piecewise smooth sets*. We later prove that such partitions are dense among partitions of D by sets of finite perimeters.²

Constructing a recovery sequence for \mathcal{U} induced by piecewise smooth sets. Let $Y_r^n = A_r \cap X_n$ denote the restriction of A_r to the first n data points. Now, let us consider the transportation maps $\{T_n\}_{n \in \mathbb{N}}$ from Proposition 4. We let A_r^n be the set for which $\mathbf{1}_{A_r^n} = \mathbf{1}_{Y_r^n} \circ T_n$.

We first notice that the fact that B_r has a piecewise smooth boundary in \mathbb{R}^d and the fact that $\|Id - T_n\|_\infty \rightarrow 0$, imply that

$$(7.7) \quad \|\mathbf{1}_{A_r^n} - \mathbf{1}_{A_r}\|_{L^1(\nu)} \leq C_0(B_r) \|Id - T_n\|_\infty,$$

where $C_0(B_r)$ denotes some constant that depends on the set B_r . This inequality follows from the formulas for the volume of tubular neighborhoods (see [44]). In particular, note that by the change of variables (2.2) we have, $|Y_r^n| = |A_r^n| \rightarrow |A_r|$ as $n \rightarrow \infty$, so that in particular we can assume that $|Y_r^n| \neq 0$. We define $u_r^n := \mathbf{1}_{Y_r^n}/|Y_r^n|$ as the corresponding normalized indicator function. We claim that $\mathcal{U}_n := (u_1^n, \dots, u_R^n)$ furnishes the desired recovery sequence.

To see that $\mathcal{U}_n \in \mathcal{M}_n(D)$ we first note that each $u_r^n \in \text{Ind}_n(D)$ by construction. On the other hand, the fact that $\{A_1, \dots, A_R\}$ forms a partition of D implies that $\{Y_1^n, \dots, Y_R^n\}$ defines a partition of X_n . As a consequence,

$$E_n(\mathcal{U}_n) = \sum_{r=1}^R GTV_{n, \varepsilon_n}(u_r^n)$$

by definition of the E_n functionals.

Using (7.7), we can proceed as in remark 5.1 in [18]. In particular, we can assume that η has the form $\eta(z) = a$ for $z < b$ and $\eta(z) = 0$ otherwise. We set $\tilde{\varepsilon}_n := \varepsilon_n + \frac{2}{b}\|Id - T_n\|_\infty$. Recall that by assumption $\|Id - T_n\|_\infty \ll \varepsilon_n$, and thus $\tilde{\varepsilon}_n$ is a perturbation of ε_n . Define the non-local total variation $\widetilde{TV}_{\tilde{\varepsilon}_n}$ of an integrable function $u \in L^1(\nu)$ as

$$\widetilde{TV}_{\tilde{\varepsilon}_n}(u) := \frac{1}{\tilde{\varepsilon}_n^{d+1}} \int_{D \times D} \eta\left(\frac{|x - y|}{\tilde{\varepsilon}_n}\right) |u(x) - u(y)| \rho(x) \rho(y) \, dx dy.$$

Using the definition of $\tilde{\varepsilon}_n$, and the form of the kernel η , we deduce that for all $n \in \mathbb{N}$, and almost every $x, y \in D$ we have

$$\eta\left(\frac{|T_n(x) - T_n(y)|}{\varepsilon_n}\right) \leq \eta\left(\frac{|x - y|}{\tilde{\varepsilon}_n}\right).$$

This inequality and a change of variables (see 2.2) implies that

$$\frac{\varepsilon_n^{d+1}}{\tilde{\varepsilon}_n^{d+1}} GTV_{n, \varepsilon_n}(\mathbf{1}_{Y_r^n}) \leq \widetilde{TV}_{\tilde{\varepsilon}_n}(\mathbf{1}_{A_r^n}).$$

A straightforward computation shows that there exists a constant K_0 such that

$$|\widetilde{TV}_{\tilde{\varepsilon}_n}(\mathbf{1}_{A_r^n}) - \widetilde{TV}_{\tilde{\varepsilon}_n}(\mathbf{1}_{A_r})| \leq \frac{K_0}{\tilde{\varepsilon}_n} \|\mathbf{1}_{A_r^n} - \mathbf{1}_{A_r}\|_{L^1(\nu)} \leq K_0 C_0(B_r) \frac{\|Id - T_n\|_\infty}{\tilde{\varepsilon}_n}.$$

²Note that unlike in the two-class case, due to "triple junctions", one cannot approximate a general partition by a partition with sets with smooth boundaries. This makes the construction more complicated.

Since $\frac{\varepsilon_n}{\varepsilon_n} \rightarrow 1$, the previous inequalities imply that

$$\limsup_{n \in \mathbb{N}} GTV_{n\varepsilon_n}(\mathbf{1}_{Y_r^n}) \leq \limsup_{n \in \mathbb{N}} \widetilde{TV}_{\varepsilon_n}(\mathbf{1}_{A_r^n}) = \limsup_{n \in \mathbb{N}} \widetilde{TV}_{\varepsilon_n}(\mathbf{1}_{A_r}).$$

Finally, from remark 4.3 in [18] we deduce that

$$\limsup_{n \rightarrow \infty} \widetilde{TV}_{\varepsilon_n}(\mathbf{1}_{A_r^n}) \leq \sigma_\eta TV(\mathbf{1}_{A_r}),$$

and thus we conclude that $\limsup_{n \rightarrow \infty} GTV_{n,\varepsilon_n}(\mathbf{1}_{A_r}) \leq \sigma_\eta TV(\mathbf{1}_{A_r})$. As a consequence we have

$$\limsup_{n \rightarrow \infty} GTV_{n,\varepsilon_n}(u_r^n) = \limsup_{n \rightarrow \infty} \frac{GTV_{n,\varepsilon_n}(\mathbf{1}_{Y_r^n})}{B_n(\mathbf{1}_{Y_r^n})} \leq \sigma_\eta \frac{TV(\mathbf{1}_{A_r})}{B(\mathbf{1}_{A_r})}$$

for each r , by continuity of the balance term. From the previous computations we conclude that $E_n(\mathcal{U}_n) \rightarrow E(\mathcal{U})$, and from 7.7, we deduce that $\mathcal{U}_n \rightarrow \mathcal{U}$ in the TL^1 -sense, so that \mathcal{U}_n does furnish the desired recovery sequence.

Density. To establish Proposition 22, we show that for any given $\mathcal{U} = (\tilde{\mathbf{1}}_{A_1}, \dots, \tilde{\mathbf{1}}_{A_R})$ where each of the sets A_r has finite perimeter, there exists a sequence $\{\mathcal{U}_m = (\tilde{\mathbf{1}}_{A_1^m}, \dots, \tilde{\mathbf{1}}_{A_R^m})\}_{m \in \mathbb{N}}$, where each of the \mathcal{U}_m is induced by piecewise smooth sets, and such that for every $r \in \{1, \dots, R\}$

$$\mathbf{1}_{A_r^m} \xrightarrow{L^1(\nu)} \mathbf{1}_{A_r},$$

and

$$\lim_{m \rightarrow \infty} TV(\mathbf{1}_{A_r^m}; D) = TV(\mathbf{1}_{A_r}; D).$$

Note that in fact, by establishing the existence of such approximating sequence, it immediately follows that $\mathcal{U}_m \rightarrow \mathcal{U}$ in $(L^1(\nu))^R$ and that $\lim_{m \rightarrow \infty} E(\mathcal{U}_m) = E(\mathcal{U})$ (by continuity of the balance terms). We provide the construction of the approximating sequence $\{\mathcal{U}_m\}_{m \in \mathbb{N}}$ through the sequence of three lemmas presented below.

Lemma 26. *Let $\{A_1, \dots, A_R\}$ denote a collection of open and bounded sets with smooth boundary in \mathbb{R}^d that satisfy*

$$(7.8) \quad \mathcal{H}^{d-1}(\partial A_r \cap \partial A_s) = 0, \quad \forall r \neq s.$$

Let D denote an open and bounded set. Then there exists a permutation $\pi : \{1, \dots, R\} \rightarrow \{1, \dots, R\}$ such that

$$TV(\mathbf{1}_{A_{\pi(r)} \setminus \bigcup_{s=r+1}^R A_{\pi(s)}}; D) \leq TV(\mathbf{1}_{A_{\pi(r)}}; D), \quad \forall r \in \{1, \dots, R\}.$$

Proof. The proof is by induction on R . **Base case:** Note that if $R = 1$ there is nothing to prove.

Inductive Step: Suppose that the result holds when considering any $R - 1$ sets as described in the statement. Let A_1, \dots, A_R be a collection of sets as in the statement. By the induction hypothesis it is enough to show that we can find $r \in \{1, \dots, R\}$ such that

$$(7.9) \quad TV(\mathbf{1}_{A_r \setminus \bigcup_{s \neq r} A_s}; D) \leq TV(\mathbf{1}_{A_r}; D).$$

To simplify notation, denote by Γ_r the set ∂A_r and define a_{rs} as the quantity

$$a_{rs} := \int_{\Gamma_r \cap (A_s \setminus \bigcup_{k \neq r, k \neq s} A_k) \cap D} \rho^2(x) \, d\mathcal{H}^{d-1}(x).$$

Hypothesis (7.8) and (3.3) imply that the equality

$$(7.10) \quad TV(\mathbf{1}_{A_r \setminus \bigcup_{s \neq r} A_s}; D) = \int_{\Gamma_r \cap (\bigcup_{k \neq r} A_k)^c \cap D} \rho^2(x) \, d\mathcal{H}^{d-1}(x) + \sum_{s: s \neq r} a_{sr}$$

holds for every $r \in \{1, \dots, R\}$, as does the inequality

$$(7.11) \quad TV(\mathbf{1}_{A_r}; D) \geq \int_{\Gamma_r \cap (\bigcup_{k \neq r} A_k)^c \cap D} \rho^2(x) \, d\mathcal{H}^{d-1} + \sum_{s: s \neq r} a_{rs}.$$

If $TV(\mathbf{1}_{A_r \setminus \bigcup_{s \neq r} A_s}; D) > TV(\mathbf{1}_{A_r}; D)$ for every r then (7.11) and (7.10) would imply that

$$\sum_{s: s \neq r} a_{sr} > \sum_{s: s \neq r} a_{rs}, \quad \forall r,$$

which after summing over r would imply

$$\sum_{r=1}^R \sum_{s: s \neq r} a_{sr} > \sum_{r=1}^R \sum_{s: s \neq r} a_{rs} = \sum_{r=1}^R \sum_{s: s \neq r} a_{sr}.$$

This would be a contradiction. Hence there exists at least one r for which (7.9) holds. \square

Lemma 27. *Let D denote an open, bounded domain in \mathbb{R}^d with Lipschitz boundary and let (B_1, \dots, B_R) denote a collection of R bounded and mutually disjoint subsets of \mathbb{R}^d that satisfy*

$$(i) \quad TV(\mathbf{1}_{B_r}; \mathbb{R}^d) < +\infty \quad , \quad (ii) \quad |D\mathbf{1}_{B_r}|_{\rho^2}(\partial D) = 0 \quad \text{and} \quad (iii) \quad D \subseteq \bigcup_{r=1}^R B_r.$$

Then there exists a sequence of mutually disjoint sets $\{A_1^m, \dots, A_R^m\}$ with piecewise smooth boundaries which cover D and satisfy

$$(7.12) \quad \mathbf{1}_{A_r^m} \rightarrow_{L^1(\mathbb{R}^d)} \mathbf{1}_{B_r} \quad \text{and} \quad \lim_{m \rightarrow \infty} TV(\mathbf{1}_{A_r^m}; D) = TV(\mathbf{1}_{B_r}; D)$$

for all $1 \leq r \leq R$.

Proof. First of all note that $TV(\mathbf{1}_{B_r}; \mathbb{R}^d)$ and $|D\mathbf{1}_{B_r}|_{\rho^2}(\partial D)$ are defined considering ρ as a function from \mathbb{R}^d into \mathbb{R} . We are using the extension considered when we introduced the weighted total variation at the beginning of subsection 3.1. Given that $\rho^2 : \mathbb{R}^d \rightarrow (0, \infty)$ is lower semi-continuous and bounded below and above by positive constants then, it belongs to the class of weights considered in [6] where the weighted total variation is studied.

Let $\{\gamma_k\}_{k \in \mathbb{N}}$ denote some sequence of positive reals converging to zero and

$$J_k(x) := \frac{1}{\gamma_k^d} J\left(\frac{|x|}{\gamma_k}\right) \quad J \geq 0, \quad J \in C_c^\infty([0, 1]), \quad \int_{\mathbb{R}^d} J(x) \, dx = 1$$

a corresponding sequence of positive, radially symmetric mollifiers. Let $D_k := \{x \in \mathbb{R}^d : \text{dist}(x, D) < \gamma_k\}$ denote the open γ_k -neighborhood of the domain D . For each $k \in \mathbb{N}$ and each B_r in the collection let

$$u_r^k := J_k * \mathbf{1}_{B_r}$$

denote a smoothed version of the characteristic function.

For any test function $\Phi \in C_c^1(D; \mathbb{R}^d)$ with $|\Phi(x)| \leq \rho^2(x)$, we have

$$\int_D u_r^k \, \text{div}(\Phi(x)) \, dx = - \int_{D_k} \mathbf{1}_{B_r} \, \text{div}(J_k * \Phi(y)) \, dy \leq |D\mathbf{1}_{B_r}|_{\rho^2}(D_k).$$

The equality follows from the symmetry of J_k and the fact that $J_k * \Phi$ has support within D_k while the inequality follows from the fact that $|J_k * \Phi| \leq \rho^2$ so it produces an admissible test function in the definition of the total variation. As a consequence,

$$\limsup_{k \rightarrow \infty} TV(u_r^k; D) \leq \limsup_{k \rightarrow \infty} |D\mathbf{1}_{B_r}|_{\rho^2}(D_k) = |D\mathbf{1}_{B_r}|_{\rho^2}(\overline{D}) = |D\mathbf{1}_{B_r}|_{\rho^2}(D)$$

due to the second assumption in the statement of the lemma. The fact that $u_r^k \rightarrow_{L^1(\mathbb{R}^d)} \mathbf{1}_{B_r}$ combines with the lower-semicontinuity of the total variation to imply

$$TV(\mathbf{1}_{B_r}; D) \leq \liminf_{k \rightarrow \infty} TV(u_r^k; D) \leq \limsup_{k \rightarrow \infty} TV(u_r^k; D) \leq TV(\mathbf{1}_{B_r}; D).$$

In other words, these sequences satisfy

$$u_r^k \rightarrow_{L^1(\mathbb{R}^d)} \mathbf{1}_{B_r}, \quad TV(u_r^k; D) \rightarrow TV(\mathbf{1}_{B_r}; D), \quad 0 \leq u_r^k(x) \leq 1 \quad \forall x \in \mathbb{R}^d.$$

The (u_1^k, \dots, u_R^k) also satisfy one additional property that will prove useful: there exists a constant $\alpha > 0$ so that

$$\Sigma^k(x) := \sum_{r=1}^R u_r^k(x) \geq \alpha > 0 \quad \text{for all } x \in D.$$

To see this, note that the fact that D is an open and bounded set with Lipschitz boundary implies that there exists a cone $C \subseteq \mathbb{R}^d$ with non-empty interior, a family of rotations $\{R_x\}_{x \in D}$ and $\zeta > 0$ such that for every $x \in D$,

$$x + R_x(C \cap B(0, \zeta)) \subseteq D.$$

The fact that J is radially symmetric then implies that for every $x \in D$,

$$\begin{aligned} \int_D J_k(x-y) dy &\geq \int_{x+R_x(C \cap B(0, \zeta))} J_k(x-y) dy = \\ &\int_{C \cap B(0, \frac{\zeta}{\gamma_k})} J_k(y) dy = \int_{C \cap B(0, \frac{\zeta}{\gamma_k})} J(y) dy \geq \alpha > 0 \end{aligned}$$

for some positive constant α . The summation $\Sigma^k(x)$ of all u_r^k therefore satisfies the pointwise estimate

$$\Sigma^k(x) := \sum_{r=1}^R u_r^k(x) = \int_{\mathbb{R}^d} J_k(x-y) \sum_{r=1}^R \mathbf{1}_{B_r}(y) dy \geq \int_D J_k(x-y) dy \geq \alpha$$

for all $x \in D$ as claimed.

Step 1: Now, for each u_r^k and each $t \in (0, 1)$ consider the superlevel set

$$B_r^k(t) := \{u_r^k > t\}.$$

The first claim is that, for any fixed t in $(0, 1)$, the characteristic function $\mathbf{1}_{B_r^k(t)}$ converges in $L^1(\mathbb{R}^d)$ to the characteristic function of the original set. To see this, note that

$$B_r^k(t) \setminus B_r \subset \{|u_r^k - \mathbf{1}_{B_r}| \geq t\}.$$

By Chebyshev's/Markov's inequality, if \mathcal{L}_d denotes Lebesgue measure in \mathbb{R}^d then

$$\mathcal{L}_d(B_r^k(t) \setminus B_r) \leq \mathcal{L}_d\left(\{|u_r^k - \mathbf{1}_{B_r}| > t\}\right) \leq \frac{1}{t} \|u_r^k - \mathbf{1}_{B_r}\|_{L^1} \rightarrow 0.$$

In a similar fashion,

$$\mathcal{L}_d(B_r \setminus B_r^k(t)) \leq \mathcal{L}_d\left(\{|u_r^k - \mathbf{1}_{B_r}| \geq (1-t)\}\right) \leq \frac{1}{1-t} \|u_r^k - \mathbf{1}_{B_r}\|_{L^1} \rightarrow 0.$$

As a consequence, it follows that

$$\int_{\mathbb{R}^d} |\mathbf{1}_{B_r^k(t)} - \mathbf{1}_{B_r}| dx = \mathcal{L}_d(B_r^k(t) \setminus B_r) + \mathcal{L}_d(B_r \setminus B_r^k(t)) \rightarrow 0$$

as claimed.

Step 2: The next claim is that there exists a set $\mathcal{T} \subset (0, 1)$ of full Lebesgue measure with the following property: if $t \in \mathcal{T}$ then $B_r^k(t)$ has a smooth boundary for all k and all sets $(B_1^k(t), \dots, B_R^k(t))$ in the collection. To see this, note Sard's lemma (see for example [29]) implies that for any fixed $k \in \mathbb{N}$ the set $B_r^k(t)$ has smooth boundary up to an exceptional set $\mathcal{T}_{k,r} \subset (0, 1)$ of Lebesgue measure zero. Now define the set \mathcal{T} as

$$\mathcal{T} = (0, 1) \setminus \bigcup_{k=1}^{\infty} \bigcup_{r=1}^R \mathcal{T}_{k,r}.$$

Note that \mathcal{T} has full measure since a countable union of Lebesgue-null sets has measure zero. If $t \in \mathcal{T}$ then it does not lie in any of the exceptional sets, meaning that for each k and each r the set $B_r^k(t)$ has a smooth boundary.

Step 3: We use a diagonal argument to construct an approximating sequence of partitions that are not necessarily disjoint, but satisfy the hypotheses of Lemma (7.8).

For the set B_1 , Step 1 and lower semi-continuity of the total variation imply that for all $t \in (0, 1)$

$$TV(\mathbf{1}_{B_1}; D) \leq \liminf_{k \rightarrow \infty} TV(\mathbf{1}_{B_1^k(t)}; D).$$

On the other hand, Fatou's lemma combines with the co-area formula to imply

$$\int_0^1 \liminf_{k \rightarrow \infty} TV(\mathbf{1}_{B_1^k(t)}; D) dt \leq \lim_{k \rightarrow \infty} \int_0^1 TV(\mathbf{1}_{B_1^k(t)}; D) dt = \lim_{k \rightarrow \infty} TV(u_1^k; D) = TV(\mathbf{1}_{B_1}; D).$$

In other words,

$$TV(\mathbf{1}_{B_1}; D) \leq \liminf_{k \rightarrow \infty} TV(\mathbf{1}_{B_1^k(t)}; D) \quad \text{and} \quad \int_0^1 \liminf_{k \rightarrow \infty} TV(\mathbf{1}_{B_1^k(t)}; D) dt = TV(\mathbf{1}_{B_1}; D),$$

which implies

$$\liminf_{k \rightarrow \infty} TV(\mathbf{1}_{B_1^k(t)}; D) = TV(\mathbf{1}_{B_1}; D)$$

almost everywhere. In particular, there exists a $t_1 \in \mathcal{T}$ with $0 < t_1 < \alpha/R$ and a subsequence $\{k_m\}_{m \in \mathbb{N}}$ with the property that

$$(7.13) \quad \partial B_1^{k_m}(t_1) \text{ is smooth } \forall m, \quad \lim_{m \rightarrow \infty} TV(\mathbf{1}_{B_1^{k_m}(t_1)}; D) = TV(\mathbf{1}_{B_1}; D), \quad \mathbf{1}_{B_1^{k_m}(t_1)} \xrightarrow{L^1(\nu)} \mathbf{1}_{B_1}.$$

We now pass to the set B_2 . As $\partial B_1^{k_m}(t_1)$ is smooth and bounded for all m , it has zero Lebesgue measure for all m in particular. As $u_2^{k_m}$ is smooth, lemma 2.95 in [1] implies that

$$\mathcal{H}^{d-1}(\partial B_1^{k_m}(t_1) \cap \partial B_2^{k_m}(t)) = 0$$

for almost every $t \in (0, 1)$. Let $\mathcal{T}_{2,m}$ denote the m^{th} exceptional set for which this property does not hold. Define the set

$$\mathcal{T}_2 := \mathcal{T} \setminus \bigcup_{m=1}^{\infty} \mathcal{T}_{2,m},$$

which has full Lebesgue measure. By definition, if $t \in \mathcal{T}_2$ then $\partial B_2^{k_m}(t)$ is smooth for all m and

$$\mathcal{H}^{d-1}(\partial B_1^{k_m}(t_1) \cap \partial B_2^{k_m}(t)) = 0$$

for all m as well. Along the subsequence $\{k_m\}$, the lower semi-continuity property still holds,

$$TV(\mathbf{1}_{B_2}; D) \leq \liminf_{k \rightarrow \infty} TV(\mathbf{1}_{B_2^{k_m}(t)}; D),$$

as does the argument based on Fatou's lemma and the co-area formula. In particular, there exists a further subsequence $\{k_{m_l}\}_{l \in \mathbb{N}}$ and a $t_2 \in \mathcal{T}_2$ with $0 < t_2 < \alpha/R$ so that (7.13) holds along this subsequence. The analogous properties hold for the sets $\{B_2^{k_{m_l}}(t_2)\}$ as well. Moreover, the relation

$$\mathcal{H}^{d-1}(\partial B_1^{k_{m_l}}(t_1) \cap \partial B_2^{k_{m_l}}(t_2)) = 0$$

also holds along this subsequence. By extracting $(R-2)$ more subsequences in this way, we obtain a subsequence that we denote simply by k_m of the original sequence together with a sequence of sets $B_r^{k_m}(t_r)$ with $0 < t_r < \alpha/R$ that satisfy

$$(7.14) \quad \begin{aligned} &\partial B_r^{k_m}(t_r) \text{ is smooth } \forall m, \quad \lim_{m \rightarrow \infty} TV(\mathbf{1}_{B_r^{k_m}(t_r)}; D) = TV(\mathbf{1}_{B_r}; D), \quad \mathbf{1}_{B_r^{k_m}(t_r)} \xrightarrow{L^1(\nu)} \mathbf{1}_{B_r}, \\ &\mathcal{H}^{d-1}(\partial B_r^{k_m}(t_r) \cap \partial B_s^{k_m}(t_s)) = 0 \end{aligned}$$

for all m and all $r \neq s$.

Step 4: We now use the sets constructed in the previous step and lemma 26 to complete the proof. Let $B_r^m := B_r^{k_m}(t_r)$. We claim that the sets (B_1^m, \dots, B_R^m) cover D as well. To see this, suppose there exists

$$x \in D \setminus \left(\bigcup_{r=1}^R B_r^m \right).$$

This would imply that $u_r^{k_m}(x) \leq t_r$ for all r by definition. In turn,

$$\Sigma^{k_m}(x) \leq \sum_{r=1}^R t_r < \alpha,$$

which contradicts the estimate on Σ^{k_m} obtained earlier. Due to (7.14) and Lemma 26, for each $m \in \mathbb{N}$ there exists a permutation $\pi_m : \{1, \dots, R\} \rightarrow \{1, \dots, R\}$ with the property that

$$TV(\mathbf{1}_{A_r^m}; D) \leq TV(\mathbf{1}_{B_{\pi_m(r)}^m}; D)$$

for all $1 \leq r \leq R$, where A_r^m denotes the set

$$A_r^m := B_r^m \setminus \bigcup_{s=\pi_m^{-1}(r)+1}^R B_s^m.$$

Each A_r^m has a piecewise smooth boundary for all m due to the fact that each B_r^m has a smooth boundary. The disjointness of (B_1, \dots, B_R) combines with the L^1 -convergence of $\mathbf{1}_{B_r^m}$ to $\mathbf{1}_{B_r}$ to show that

$$\mathbf{1}_{A_r^m} \rightarrow_{L^1(\mathbb{R}^d)} \mathbf{1}_{B_r}$$

as well. This combines with lower semi-continuity of the total variation to imply

$$\begin{aligned} TV(\mathbf{1}_{B_r}; D) &\leq \liminf_{m \rightarrow \infty} TV(\mathbf{1}_{A_r^m}; D) \\ &\leq \limsup_{m \rightarrow \infty} TV(\mathbf{1}_{A_r^m}; D) \leq \limsup_{m \rightarrow \infty} TV(\mathbf{1}_{B_{\pi_m(r)}^m}; D) = TV(\mathbf{1}_{B_r}; D). \end{aligned}$$

Finally, noting that

$$D \subset \bigcup_{r=1}^R B_r^m = \bigcup_{r=1}^R A_r^m$$

and that the A_r^m are pairwise disjoint yields the claim. \square

To complete the construction, and therefore to conclude the proof of theorem 2, we need to verify the hypotheses (i – ii) of the previous lemma. This is the content of our final lemma.

Lemma 28. *Let D be an open and bounded domain with Lipschitz boundary and let $\{A_1, \dots, A_R\}$ denote a disjoint collection of sets that satisfy*

$$A_r \subset D \quad \text{and} \quad TV(\mathbf{1}_{A_r}; D) < \infty.$$

Then, there exists a disjoint collection of bounded sets (B_1, \dots, B_R) that satisfy $B_r \cap D = A_r$ together with the properties

$$(i) \quad TV(\mathbf{1}_{B_r}; \mathbb{R}^d) < +\infty \quad \text{and} \quad (ii) \quad |D\mathbf{1}_{B_r}|_{\rho^2}(\partial D) = 0.$$

The proof follows from remark 3.43 in [1] (which with minimal modifications applies to total variation with weight ρ^2). \square

8. NUMERICAL EXPERIMENTS

We now present numerical experiments to provide a concrete demonstration and visualization of the theoretical results developed in this paper. These experiments focus on elucidating when and how minimizers of the graph-based Cheeger Cut problem,

$$(8.1) \quad u_n^* \in \operatorname{argmin}_{u \in L^1(\nu_n)} E_n(u) \quad \text{with} \quad B_n(u) := \min_{c \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |u(\mathbf{x}_i) - c|,$$

converge in the appropriate sense to a minimizer of the continuum Cheeger Cut problem

$$(8.2) \quad u^* \in \operatorname{argmin}_{u \in L^1(\nu)} E(u) \quad \text{with} \quad B(u) := \min_{c \in \mathbb{R}} \int_D |u(x) - c| \, dx.$$

We always take $\rho(x) := 1/\operatorname{vol}(D)$ as the constant density. The data points $X_n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ therefore represent i.i.d. samples from the uniform distribution. We consider the following two rectangular domains

$$D_1 := (0, 1) \times (0, 4) \quad \text{and} \quad D_2 := (0, 1) \times (0, 1.5)$$

in our experiments. We may easily compute the optimal continuum Cheeger Cut for these domains. The characteristic function

$$\mathbf{1}_{A_1}(x) \quad \text{for} \quad A_1 := \{(x, y) \in D_1 : y > 2\},$$

when appropriately normalized, provides a minimizer $u_1^* \in L^1(\nu)$ of the continuum Cheeger Cut in the former case, while the characteristic function

$$\mathbf{1}_{A_2}(x) \quad \text{for} \quad A_2 := \{(x, y) \in D_2 : y > 0.75\}$$

analogously furnishes a minimizer $u_2^* \in L^1(\nu)$ in the latter case. Figure 2 provides an illustration of a sequence of discrete partitions, computed from the graph-based Cheeger Cut problem, to the optimal continuum Cheeger Cut.

Each of our experiments utilize the nearest neighbor kernel $\eta(z) = \mathbf{1}_{\{|z| \leq 1\}}$ for the computation of the similarity weights,

$$w_{i,j} = \mathbf{1}_{\{\|\mathbf{x}_i - \mathbf{x}_j\| \leq \varepsilon_n\}},$$

so that the graphs $\mathcal{G}_n = (X_n, W_n)$ correspond to ε_n -nearest neighbor graphs (a.k.a. random geometric graphs). We use the steepest descent algorithm from [12] to solve the graph-based Cheeger Cut problem on these graphs. We initialized the algorithm with the “ground-truth” partition $Y_n^i := A_i \cap X_n$, and we terminated the algorithm once three consecutive iterates show 0% change in the corresponding partition of the graph. We let Y_n^* denote the partition of \mathcal{G}_n returned by the algorithm, which we view as the “optimal” solution of the graph-based Cheeger Cut problem. We quantify the error between the optimal

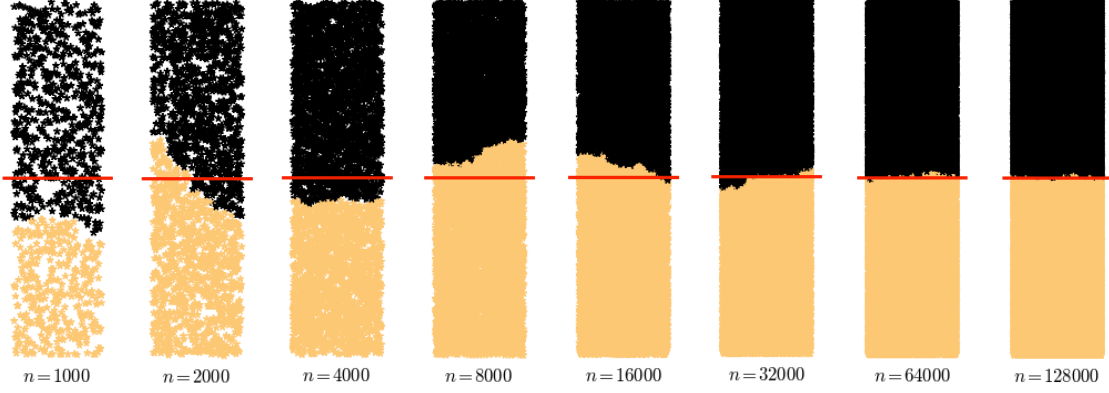


FIGURE 2. Visualization of the convergence process. Each figure depicts a computed optimal partition Y_n^* (in black) of one random realization of the random geometric graph $\mathcal{G}_n = (X_n, W_n)$ for each $k \in \{0, 1, \dots, 7\}$, where $n = 1000 \times 2^k$, $\varepsilon = n^{-0.3}$ and the domain considered is D_1 . Note that the scaling of ε with respect to n falls within the context of our theoretical results. The red line indicates the optimal cut, i.e. the boundary of the set $A_1 := \{(x, y) \in D_1 : y > 2\}$, at the continuum level.

continuum partition $A_i \subsetneq D_i$ and the n^{th} optimal graph-based partition Y_n^* of \mathcal{G}_n simply by using the percentage of misclassified data points, i.e.

$$(8.3) \quad e_n = \frac{1}{n} \sum_{i=1}^n |\mathbf{1}_{Y_n^i}(\mathbf{x}_i) - \mathbf{1}_{Y_n^*}(\mathbf{x}_i)|.$$

If $T_n(x)$ denotes a sequence of transportation maps between ν_n and ν that satisfy $\|Id - T_n\|_\infty = o(1)$, then by the change of variables (2.2)

$$e_n = \int_D |\mathbf{1}_{A_i} \circ T_n(x) - \mathbf{1}_{Y_n^*} \circ T_n(x)| \, dx.$$

By the triangle inequality, we therefore obtain

$$\begin{aligned} \|\mathbf{1}_{A_i} - \mathbf{1}_{Y_n^*} \circ T_n\|_{L^1(\nu)} &:= \int_D |\mathbf{1}_{A_i}(x) - \mathbf{1}_{Y_n^*} \circ T_n(x)| \, dx \\ &\leq e_n + \int_D |\mathbf{1}_{A_i}(x) - \mathbf{1}_{A_i} \circ T_n(x)| \, dx \leq e_n + O(\|Id - T_n\|_\infty). \end{aligned}$$

The last inequality follows since each A_i has a piecewise smooth boundary. In this way, if $\|Id - T_n\|_\infty = o(1)$ then verifying $e_n = o(1)$ suffices to show that TL^1 convergence of minimizers holds. Under this assumption, i.e. $\|Id - T_n\|_\infty = o(1)$, a similar argument shows that having $e_n = o(1)$ is equivalent to TL^1 convergence in the context of our experiments.

To check convergence, and to explore the issues related to Remark (2), we perform exhaustive numerical experiments for three distinct scalings of ε_n with respect to the total number of sample points on the domain D_2 . Specifically, we consider the scalings

$$\varepsilon_n = n^{-0.3}, \quad \varepsilon_n = 2 \left(\frac{\log n}{\pi n} \right)^{1/2}, \quad \text{and} \quad \varepsilon_n = \left(\frac{\log n}{\pi n} \right)^{1/2}.$$

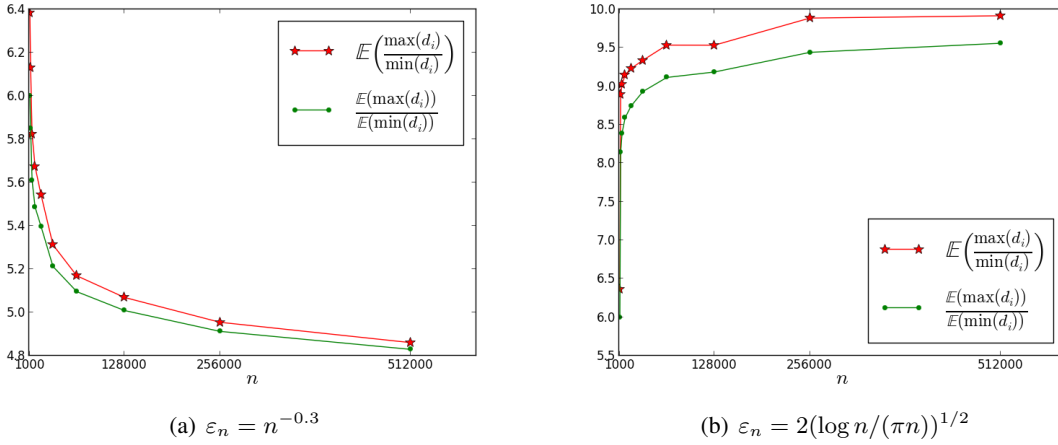


FIGURE 3. Graph regularity. We work with the domain D_2 . For each scaling of ε_n with n , the corresponding plot depicts two measures of regularity for the sequence of random geometric graphs. The first measure (in red) is the average $\mathbb{E}(\max(d_i)/\min(d_i))$, i.e. the average ratio of the maximal degree $\max(d_i)$ of \mathcal{G}_n to the minimal degree. For each n , the average is computed over 1,440 independent graph realizations. The second measure (in green) corresponds to the ratio of the average maximal degree to the average minimal degree, computed over 1,440 independent trials as before. The graphs with $\varepsilon_n = n^{-0.3}$ become increasingly regular while the graphs with $\varepsilon_n = 2(\log n / (\pi n))^{1/2}$ become increasingly irregular.

These scalings correspond to three distinct types of random geometric graphs. The first scaling falls well within the acceptable bounds for ε_n covered by our consistency theorems. In particular, \mathcal{G}_n is almost surely connected in this regime. The second scaling also gives rise to a sequence \mathcal{G}_n of connected random geometric graphs (see [22], [32]). However, the geometric graphs \mathcal{G}_n exhibit rather different structural properties in this case; if $\varepsilon_n = n^{-0.3}$ then the graphs \mathcal{G}_n become increasingly regular as $n \rightarrow \infty$, while if $\pi\varepsilon_n^2 = 2(\log n)/n$ then the graphs \mathcal{G}_n become increasingly irregular. See Figure 3 for an illustration. The final scaling corresponds to a scaling below the connectivity threshold of random geometric graphs (c.f. [22], [32]). The graphs \mathcal{G}_n are almost surely disconnected under this scaling. However, in this regime each \mathcal{G}_n has a “giant component” (i.e. a connected subgraph \mathcal{H}_n of \mathcal{G}_n) that contains all but a small handful of vertices (c.f. Figure 5 at left).

We designed our experiments to explore the extent to which connectivity, and connectivity alone, is responsible for consistency of balanced cuts. The first scaling $\varepsilon_n = n^{-0.3}$ serves as a benchmark or control. It falls within the context of our consistency theorems, and so provides a means of determining the “typical” behavior of balanced cut algorithms when consistency holds. The second scaling, which falls outside the realm of our consistency results, tests whether connected graphs with different structural properties still lead to consistent results. The final scaling probes the realm where connectivity fails, but in a mild and easily correctable way. As the theory outlined above indicates, if we pose the balance cut minimization over the full graph \mathcal{G}_n then we can no longer expect consistency to hold. These graphs pose no practical difficulty, however, as we may simply extract the giant component \mathcal{H}_n of each \mathcal{G}_n and then minimize the balanced cut over this connected subgraph. We simply assign each vertex in $\mathcal{G}_n \setminus \mathcal{H}_n$ to one of the two classes uniformly at random. Our last experiment explores whether consistency might still hold using this modified approach.

$n =$	$1k$	$2k$	$4k$	$8k$	$16k$	$32k$	$64k$	$128k$
$\varepsilon_n = n^{-0.3} :$								
$\mathbb{E}(e_n)$.0778	.0609	.0471	.0397	.0321	.0238	.0205	.0161
Trials	20063	4800	1200	1536	1536	1008	192	156
$\varepsilon_n = 2(\log n/(\pi n))^{1/2} :$								
$\mathbb{E}(e_n)$.0717	.0605	.0507	.0431	.0366	.0309	\times	\times
Trials	10080	10080	4032	1008	1008	304	\times	\times
$\varepsilon_n = (\log n/(\pi n))^{1/2} :$								
$\mathbb{E}(e_n)$.3243	.1977	.1203	.0891	.0672	.0545	.0442	\times
Trials	2896	16128	8064	2016	1008	592	80	\times

TABLE 1. Average error $\mathbb{E}(e_n)$ between partitions. For each n and each scaling of ε_n , we obtained an estimate of the average error $\mathbb{E}(e_n)$ by computing the mean of (8.3) over the indicated number of independent trials. Figure 4 provides a corresponding error plot.

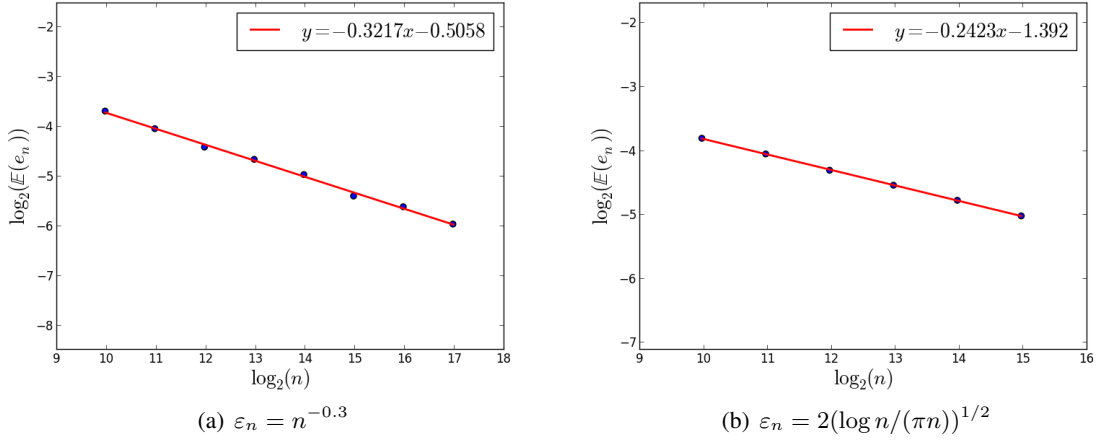


FIGURE 4. Log-log plot of the expected errors computed in Table 1, together with a corresponding linear approximation.

Table 1, Figure 4 and Figure 5 report the results of these experiments. In all cases, we measure error by using the expected number of misclassified points (8.3) averaged over the number of trials indicated in Table 1. We used a smaller number of trials for large n (or no trial at all, indicated by a \times) simply due to the overwhelming computational burden. The measure of error considered in these experiments, taken alone, does not suffice to show convergence in the almost sure sense as provided by our consistency theorems. It does, however, indicate consistency in the weaker sense of convergence in probability (via Markov's inequality). The algorithm we use to optimize the discrete Cheeger cut also relies upon a non-convex minimization [12], so we cannot say with certainty that the corresponding computed optimizers are global. Instead, initializing the algorithm with the “ground truth” partition

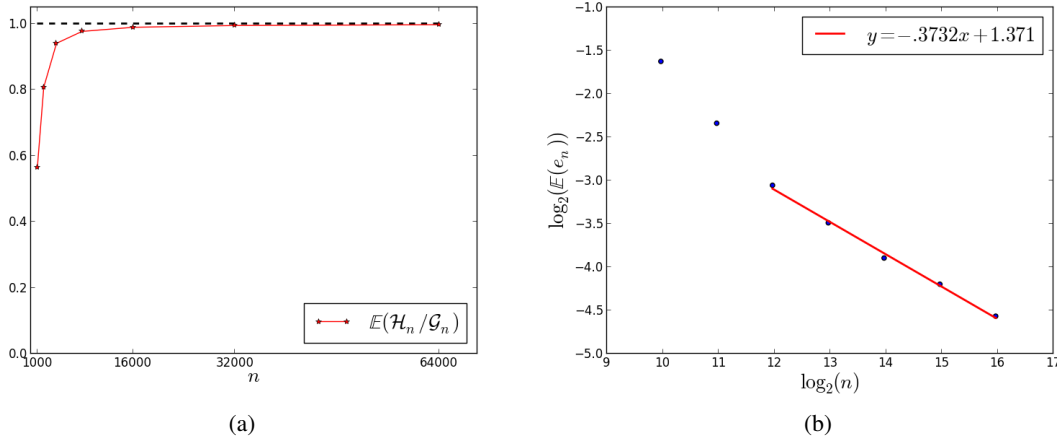


FIGURE 5. Convergence in the disconnected regime $\varepsilon_n = (\log n/(\pi n))^{1/2}$. At left: the expected fraction of vertices that lie in the giant component \mathcal{H}_n of the disconnected random geometric graph \mathcal{G}_n . At right: a log-log plot of the expected errors computed in Table 1, together with a corresponding linear approximation for n large.

biases the algorithm toward the correct cut. If the algorithm were to fail under these circumstances, it would provide strong numerical evidence *against* consistency.

The results appear rather similar regardless of whether ε_n lies in the strongly connected ($\varepsilon_n = n^{-0.3}$), weakly connected ($\varepsilon_n = 2(\log n/(\pi n))^{1/2}$) or weakly disconnected ($\varepsilon_n = (\log n/(\pi n))^{1/2}$) regimes. Indeed, in each case the error $\mathcal{E}(e_n)$ decays to zero with a polynomial rate. In other words, the varying structural properties of the random geometric graphs in these regimes do not seem to play much of a role. While certainly not conclusive evidence, it seems reasonable to conjecture that consistency should hold, perhaps in the weaker probabilistic sense, for ε_n as small as the critical scaling for connectivity. We leave a further exploration of this for future research.

Acknowledgements. The authors are grateful to ICERM, where part of the research was done during the research cluster *Geometric analysis methods for graph algorithms*. DS and NGT are grateful to NSF (grant DMS-1211760) for its support. JvB was supported by NSF grant DMS 1312344. TL was supported by NSF (grant DMS-1414396). The authors would like to thank the Center for Nonlinear Analysis of the Carnegie Mellon University for its support.

REFERENCES

- [1] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.
- [2] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Proceedings of the 47th Annual Symposium on Foundations of Computer Science (FOCS '06)*, pages 475–486, 2006.
- [3] E. Arias-Castro and B. Pelletier. On the convergence of maximum variance unfolding. *The Journal of Machine Learning Research*, 14(1):1747–1770, 2013.
- [4] E. Arias-Castro, B. Pelletier, and P. Pudlo. The normalized graph cut and Cheeger constant: from discrete to continuous. *Advances in Applied Probability*, 44:907–937, 2012.
- [5] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009.
- [6] A. Baldi. Weighted BV functions. *Houston J. Math.*, 27(3):683–705, 2001.
- [7] M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. *Advances in Neural Information Processing Systems (NIPS)*, 19:129, 2006.

- [8] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *J. Comput. System Sci.*, 74(8):1289–1308, 2008.
- [9] G. Bellettini, G. Bouchitté, and I. Fragalà. BV functions with respect to a measure and relaxation of metric integral functionals. *J. Convex Anal.*, 6(2):349–366, 1999.
- [10] A. Braides. *Gamma-Convergence for Beginners*. Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, 2002.
- [11] X. Bresson and T. Laurent. Asymmetric Cheeger cut and application to multi-class unsupervised clustering. CAM report 12-27, UCLA, 2012.
- [12] X. Bresson, T. Laurent, D. Uminsky, and J. von Brecht. Convergence and energy landscape for Cheeger cut clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1394–1402, 2012.
- [13] X. Bresson, T. Laurent, D. Uminsky, and J. von Brecht. Multiclass total variation clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.
- [14] J. Cheeger. A Lower Bound for the Smallest Eigenvalue of the Laplacian. *Problems in Analysis*, pages 195–199, 1970.
- [15] F. R. K. Chung. *Spectral Graph Theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC, 1997.
- [16] G. Dal Maso. *An Introduction to Γ -convergence*. Springer, 1993.
- [17] R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.
- [18] N. García Trillos and D. Slepčev. Continuum limit of total variation on point clouds. *Preprint*, 2014.
- [19] N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞ -transportation distance. *Preprint*, 2014.
- [20] E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.
- [21] A. Goel, S. Rai, and B. Krishnamachari. Sharp thresholds for monotone properties in random geometric graphs. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 580–586, New York, 2004. ACM.
- [22] P. Gupta and P. R. Kumar. Critical power for asymptotic connectivity in wireless networks. In *Stochastic analysis, control, optimization and applications*, Systems Control Found. Appl., pages 547–566. Birkhäuser Boston, Boston, MA, 1999.
- [23] L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design*, 11:1074–1085, 1992.
- [24] J. Hartigan. Consistency of single linkage for high density clusters. *J. Amer. Statist. Assoc.*, 76:388–394., 1981.
- [25] M. Hein, J.-Y. Audibert, and U. Von Luxburg. From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians. In *Learning theory*, pages 470–485. Springer, 2005.
- [26] M. Hein and T. Bühler. An Inverse Power Method for Nonlinear Eigenproblems with Applications in 1-Spectral Clustering and Sparse PCA. In *Advances in Neural Information Processing Systems (NIPS)*, pages 847–855, 2010.
- [27] M. Hein and S. Setzer. Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.
- [28] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.
- [29] G. Leoni. *A first course in Sobolev spaces*, volume 105 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2009.
- [30] M. Maier, U. von Luxburg, and M. Hein. How the result of graph clustering methods depends on the construction of the graph. *ESAIM: Probability and Statistics*, 17:370–418, 1 2013.
- [31] M. Penrose. A strong law for the longest edge of the minimal spanning tree. *Ann. Probab.*, 27(1):246–260, 1999.
- [32] M. Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003.
- [33] D. Pollard. Strong consistency of k-means clustering. *ann. statist.* 9 135–140. *Annals of Statistics*, 9:135–140, 1981.
- [34] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):888–905, 2000.
- [35] A. Singer. From graph to manifold Laplacian: the convergence rate. *Appl. Comput. Harmon. Anal.*, 21(1):128–134, 2006.
- [36] D. A. Spielman and S. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 81–90, 2004.
- [37] D. A. Spielman and S. Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013.

- [38] A. Szlam and X. Bresson. Total variation and Cheeger cuts. In *International Conference on Machine Learning (ICML)*, pages 1039–1046, 2010.
- [39] D. Ting, L. Huang, and M. I. Jordan. An analysis of the convergence of graph Laplacians. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.
- [40] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
- [41] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. Technical Report TR 134, Max Planck Institute for Biological Cybernetics, 2004.
- [42] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.
- [43] Y.-C. Wei and C.-K. Cheng. Towards efficient hierarchical designs by ratio cut partitioning. In *Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers., 1989 IEEE International Conference on*, pages 298–301. IEEE, 1989.
- [44] H. Weyl. On the Volume of Tubes. *Amer. J. Math.*, 61(2):461–472, 1939.
- [45] S. X. Yu and J. Shi. Multiclass spectral clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 313–319. IEEE, 2003.

¹ DEPARTMENT OF MATHEMATICAL SCIENCES, CARNEGIE MELLON UNIVERSITY, PITTSBURGH, PA, 15213, USA., TEL. +412 268-2545, EMAILS: NGARCIAT@ANDREW.CMU.EDU, SLEPCEV@MATH.CMU.EDU

² DEPARTMENT OF MATHEMATICS AND STATISTICS, CALIFORNIA STATE UNIVERSITY, LONG BEACH LONG BEACH, CA 90840, USA., EMAIL: JAMES.VONBRECHT@CSULB.EDU

³ DEPARTMENT OF MATHEMATICS, LOYOLA MARYMOUNT UNIVERSITY, 1 LMU DR, LOS ANGELES, CA 90045, USA., EMAIL: THOMAS.LAURENT@LMU.EDU

⁴ INSTITUTE OF ELECTRICAL ENGINEERING, SWISS FEDERAL INSTITUTE OF TECHNOLOGY (EPFL), 1015 LAUSANNE, SWITZERLAND, EMAIL: XAVIER.BRESSON@EPFL.CH