Journal of Machine Learning Research 17 (2016) 1-46 Submitted 11/14; Revised 2/16; Published 10/16

Consistency of Cheeger and Ratio Graph Cuts

Nicolás García Trillos ngarciat@andrew.cmu.edu
Dejan Slepčev slepcev@math.cmu.edu
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15213, USA

James von Brecht James.vonBrecht@csulb.edu
Department of Mathematics and Statistics
California State University, Long Beach
Long Beach, CA 90840, USA

Thomas Laurent Thomas.Laurent@lmu.edu
Department of Mathematics
Loyola Marymount University
1 LMU Dr
Los Angeles, CA 90045, USA

Xavier Bresson xavier.bresson@epfl.ch
Institute of Electrical Engineering
Swiss Federal Institute of Technology (EPFL)
1015 Lausanne, Switzerland

Editor: Matthias Hein

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.

Keywords: data clustering, balanced cut, consistency, graph partitioning

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 the approaches to be robust to outliers lead to the introduction of "balance" terms and objective functionals such as Cheeger cut and closely related edge expansion (Arora et al., 2009; Bresson and Laurent, 2012; Bresson et al., 2012; Kannan et al., 2004; Szlam and Bresson, 2010), ratio cut (Hagen and Kahng, 1992; Hein and Setzer, 2011; von Luxburg, 2007; Wei and Cheng, 1989), normalized cut (Arias-Castro et al., 2012; Shi and Malik, 2000; von Luxburg, 2007), and conductance (sparsest cut) (Arora et al., 2009; Kannan et al., 2004; Spielman and Teng, 2004). Such functionals have been extended by Bresson et al. (2013); Yu and Shi (2003) to treat multiclass partitioning. The balanced cuts above have been widely studied theoretically and used computationally. The algorithms of Andersen et al. (2006); Spielman and Teng (2004, 2013) use local clustering algorithms to compute balanced cuts of large graphs. Total variation based algorithms (Bresson et al., 2012, 2013; Hein and Bühler, 2010; Hein and Setzer, 2011; Szlam and Bresson, 2010) are also used to optimize either the conductance or the edge expansion of a graph. Closely related are the spectral approaches to clustering (Shi and Malik, 2000; von Luxburg, 2007) 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, $n \to \infty$, clustering methods should exhibit *consistency*. That is, the clustering of the data X_n should converge as $n \to \infty$ toward a specific clustering of the underlying ground-truth domain. In this paper we characterize in a precise manner when and how the minimizers of a ratio, Cheeger, sparsest, and normalized 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.

An important consideration when investigating consistency of algorithms is how 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 multi-way 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 that allows one to consider a limit of a family of variational problems (Braides, 2002; Dal Maso, 1993). In the recent work of García Trillos and Slepčev (2016), 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 (1981) has proved the consistency of k-means clustering.

Consistency of k-means clustering for paths with regularization was recently studied by Thorpe et al. (2015), using a similar viewpoint to those of this paper. Consistency for a class of single linkage clustering algorithms was shown by Hartigan (1981). Arias-Castro and Pelletier (2013) have proved the consistency of low-dimensional embeddings via the maximum variance unfolding. Consistency of spectral clustering was rigorously considered by von Luxburg, Belkin, and Bousquet (2004, 2008). 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 $n \to \infty$) continuum problem being a nonlocal one. Belkin and Niyogi (2006) consider the spectral problem (Laplacian eigenmaps) and show that there exists a sequence $\varepsilon_n \to 0$ 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, Pelletier, and Pudlo (2012) who provide a rate on $\varepsilon_n \to 0$ 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 (Arias-Castro et al., 2012) 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 ε_n can go to zero as $n \to \infty$ for consistency to hold.

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 (2008), Giné and Koltchinskii (2006), Hein, Audibert, and Von Luxburg (2005), Narayanan, Belkin, and Niyogi (2006), Singer (2006) and Ting, Huang, and Jordan (2010). Maier, von Luxburg, and Hein (2013) 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.

A notion of convergence suitable for showing the convergence of minimizers of approximating objective functionals converge towards a minimizer of the limit functional is the notion of Γ-convergence, which was introduced by De Giorgi in the 70's and represents a standard notion of variational convergence. For detailed exposition of the properties of Γ-convergence see the books byBraides (2002) and Dal Maso (1993). Particularly relevant to our investigation are works considering nonlocal functionals converging to the perimeter or to total variation which include works by Alberti and Bellettini (1998), Savin and Valdinoci (2012), and Esedoḡlu and Otto (2015). Also related are works of Ponce (2004), who showed the Γ-convergence of nonlocal functionals related to characterization of Sobolev spaces and of Gobbino (1998) and Gobbino and Mora (2001) who investigated nonlocal approximations of the Mumford-Shah functional. In the discrete deterministic setting, works related to the Γ-convergence of functionals to continuous functionals involving perimeter include works of

Braides and Yip (2012), Chambolle, Giacomini, and Lussardi (2010), and van Gennip and Bertozzi (2012).

1.1 Graph partitioning

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

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

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

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, that is $|Y| = \frac{\sharp Y}{n}$. Well-known balance terms include

$$\text{Bal}_{\text{R}}(Y, Y^c) = 2|Y||Y^c| \qquad \text{and} \qquad \text{Bal}_{\text{C}}(Y, Y^c) = \min(|Y|, |Y^c|), \tag{1.2}$$

which correspond to ratio cut (Hagen and Kahng, 1992; Hein and Setzer, 2011; von Luxburg, 2007; Wei and Cheng, 1989) and Cheeger cut (Arora et al., 2009; Cheeger, 1970; Chung, 1997; Kannan et al., 2004) respectively[1], as well as

$$\text{Bal}_{\text{S}}(Y, Y^c) = 2\frac{\deg(Y)\deg(Y^c)}{\deg(X)^2} \qquad \text{and} \qquad \text{Bal}_{\text{N}}(Y, Y^c) = \frac{\min(\deg(Y), \deg(Y^c))}{\deg(X)}, \tag{1.3}$$

where $\deg(Y) = \sum_{i=1}^n \sum_{j \neq i}^n w_{ij}$ is the sum of weighted degrees of all vertices in Y, which correspond to sparsest cut (Arora et al., 2009; Kannan et al., 2004; Spielman and Teng, 2004) and normalized cut (Arias-Castro et al., 2012; Shi and Malik, 2000; von Luxburg, 2007) respectively. 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 one expects that generically the optimal cut is unique, since a small perturbation of the weights of a graph with a non-unique minimal balanced cut, is almost sure to lead to only one of them having the least energy).

We are also interested in multi-class balanced cuts. Specifically, in order to partition the set X into $R \ge 3$ clusters, we consider the following ratio cut functional:

$$\underset{(Y_1, \dots, Y_R)}{\text{Minimize}} \quad \sum_{r=1}^R \frac{\text{Cut}(Y_r, Y_r^c)}{|Y_r|}, \quad Y_r \cap Y_s = \emptyset \quad \text{if} \quad r \neq s, \quad \bigcup_{r=1}^R Y_r = X. \tag{1.4}$$

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

$$\text{Minimize} \quad \frac{\text{Cut}_\rho(A, A^c)}{\text{Bal}_\rho(A, A^c)}, \qquad A \subset D \ \text{ with } 0 < \nu(A) < 1. \tag{1.5}$$

1. The factor of 2 in the definition of $\text{Bal}_{\text{R}}(Y, Y^c)$ 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.

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

$$\mathrm{Cut}_\rho(A, A^c) \;:=\; \int_{\partial_D A} \rho^2(x)\, \mathrm{d}S(x). \tag{1.6}$$

We only consider cuts with weight ρ^2, since they appear as the limit of the discrete cuts we consider in this paper, as indicated in subsection 1.3.

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 (1.5) in Subsection 3.1.

If $\rho(x) = 1$ then $\mathrm{Cut}_\rho(A, A^c)$ simply corresponds to arc-length (in 2d) or surface area (in 3d). In the general case, the presence of $\rho^2(x)$ in (1.6) indicates that the regions of low density are easier to cut, so $\partial_D A$ has a tendency to pass through regions in D of low density. As in the graph case, we consider balance terms

$$\mathrm{Bal}_\rho(A, A^c) = 2|A||A^c| \qquad \text{and} \qquad \mathrm{Bal}_\rho(A, A^c) = \min(|A|, |A^c|), \tag{1.7}$$

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,

$$|A| = \nu(A) = \int_A \rho(x)\, \mathrm{d}x. \tag{1.8}$$

We also consider balance terms

$$\mathrm{Bal}_\rho(A, A^c) = 2|A|_{\rho^2}|A^c|_{\rho^2} \qquad \text{and} \qquad \mathrm{Bal}_\rho(A, A^c) = \min(|A|_{\rho^2}, |A^c|_{\rho^2}), \tag{1.9}$$

which correspond to weighted continuous equivalents of the sparsest cut and the normalized cut. Here $|A|_{\rho^2}$ stands for

$$|A|_{\rho^2} = \frac{1}{\int_D \rho^2(x)dx} \int_A \rho^2(x)\, \mathrm{d}x. \tag{1.10}$$

We refer to a pair $\{A, A^c\}$ that solves (1.5) as an *optimal balanced cut of the domain.*

The continuum equivalent of the multiway cut problem (1.4) reads

$$\underset{(A_1,\ldots,A_R)}{\text{Minimize}} \quad \sum_{r=1}^{R} \frac{\mathrm{Cut}_\rho(A_r, A_r^c)}{|A_r|}, \tag{1.11}$$

where $(A_1, \ldots, A_R)$ is an R-tuple of measurable subsets of D such that $\nu(A_r \cap A_s) = 0$ if $r \neq s$, and $\nu\left(D \setminus \bigcup_{r=1}^R A_r\right) = 0$.

1.3 Consistency of partitioning of data clouds

Let $\mathbf{x}_1, \dots, \mathbf{x}_n, \dots$ be a sequence of i.i.d random points drawn from an underlying ground-truth measure ν. Throughout the paper ν is a probability measure supported on a bounded, open set with Lipschitz boundary D. Furthermore we assume that ν has continuous density $\rho : D \to \mathbb{R}$ and that $0 < \lambda \leq \rho \leq \Lambda$ on D; in other words, ρ is bounded below and above by positive constants. We denote by $X_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, the set consisting of the first n data points.

To extract the desired information from the point cloud X_n, one builds a graph by connecting nearby points. More precisely, let $\eta : \mathbb{R}^d \to [0, \infty)$ be a radially symmetric kernel, 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. For $i, j \in \{1, \dots, n\}$, we consider the weight

$$w_{ij} = \eta\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\varepsilon}\right). \tag{1.12}$$

As more data points are available one takes smaller ε to obtain increased resolution. That is, one sets the length scale ε based on the number of available data points: $\varepsilon = \varepsilon_n$. 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 (1.5)). On Figure 1, we illustrate the partitioning of a data cloud sampled from the uniform distribution on the given domain D.

Informal statement of (a part of) the main results. *Consider $d \geq 2$ and assume the continuum balanced cut* (1.5) *has a unique minimizer $\{A, A^c\}$. Consider $\varepsilon_n > 0$ such that $\lim_{n\to\infty} \varepsilon_n = 0$ and*

$$\lim_{n\to\infty} \frac{(\log n)^{p_d}}{n^{1/d}} \frac{1}{\varepsilon_n} = 0, \tag{1.13}$$

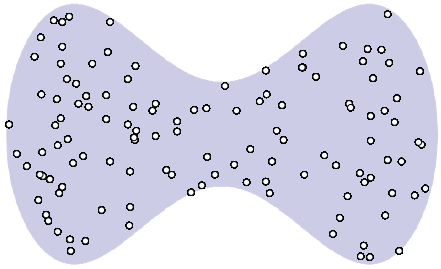
where $p_d = 1/d$ for $d \geq 3$ and $p_2 = 3/4$. Then almost surely the minimizers, $\{Y_n, Y_n^c\}$, of the balanced cut (1.1) *of the graph $\mathcal{G}_n$, converge to $\{A, A^c\}$.* Moreover, after appropriate rescaling, almost surely the minimum of problem (1.1) converges to the minimum of (1.5). *The result also holds for multiway cuts. That is, the minimizers of* (1.4) *converge towards minimizers of* (1.11)*.*

Let us make the notion of convergence of discrete partitions $\{Y_n, Y_n^c\}$ to continuum partitions $\{A, A^c\}$ precise.

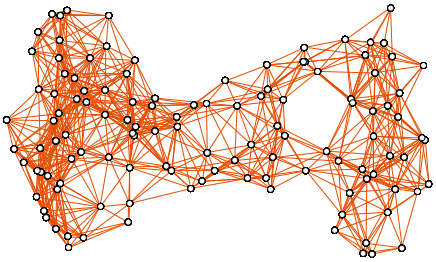
To be able to easily account for the invariance $\{Y_n, Y_n^c\} = \{Y_n^c, Y_n\}$, let $Y_{n,1} = Y_n$ and $Y_{n,2} = Y_n^c$. Let $\mathbf{1}_{Y_n,i} : X_n \to \{0, 1\}$ for $i = 1, 2$ be the characteristic function of $Y_{n,i}$ (on the set X_n). We say that $\{Y_n, Y_n^c\}$ converge towards $\{A, A^c\}$ as $n \to \infty$ if there is a sequence of indices $I : \mathbb{N} \to \{1, 2\}$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Y_{n,I(n)}}(\mathbf{x}_i) \delta_{\mathbf{x}_i} \overset{w}{\rightharpoonup} \mathbf{1}_A \, \nu \tag{1.14}$$

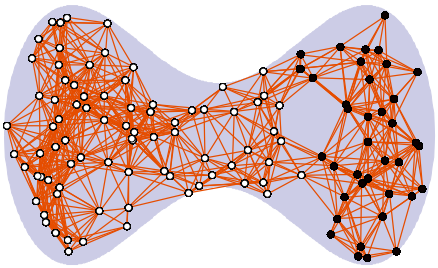
where $\overset{w}{\rightharpoonup}$ denotes the weak convergence of measures (see Dudley (2002)). Since by assumption on the points $\mathbf{x}_i$ it holds that $\frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{x}_i} \overset{w}{\rightharpoonup} \nu$, the property (1.14) is equivalent



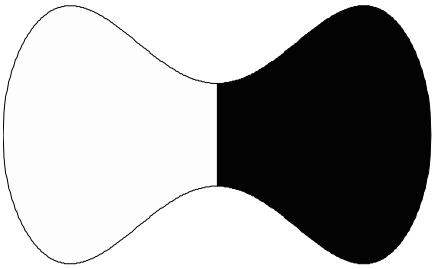
(a) A sample of $n = 120$ points.



(b) Geometric graph with $\varepsilon = 0.3$.



(c) Minimizer of Cheeger graph cut.



(d) Minimizer of continuum Cheeger cut.

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 algorithm of Bresson et al. (2012). A solution to the continuum Cheeger-cut problem is illustrated in Figure (d).

to

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{Y_{n,3-I(n)}}(\mathbf{x}_i)\delta_{\mathbf{x}_i} \xrightarrow{w} \mathbf{1}_{A^c}\,\nu$$

In Section 2 we discuss this topology in more detail and present a conceptually clearer framework, which applies to general functions (not just characteristic functions of sets).

Let us also indicate briefly why the weight ρ^2 present in the weighted perimeter (1.6) can be expected to appear in the limit of balanced graph cuts (1.1). Let $\nu_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\mathbf{x}_i}$ be the empirical measure of the sample $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Let $A \subset D$ be a set with smooth boundary and let $A_n = A \cap X_n$. Then, using $\eta_\varepsilon(z) = \eta(z/\varepsilon)/\varepsilon^d$ we get

$$\begin{aligned}
\frac{1}{n^2\varepsilon^d}\mathrm{Cut}(A_n, A_n^c) &= \frac{1}{n^2}\sum_{\mathbf{x}_i\in A_n}\sum_{\mathbf{x}_j\in A_n^c}\frac{1}{\varepsilon^d}\eta\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\varepsilon}\right)\\
&= \int_D\int_D \mathbf{1}_{A_n}(x)\mathbf{1}_{A_n^c}(y)\eta_\varepsilon\left(x-y\right)d\nu_n(x)d\nu_n(y)\\
&\sim \int_D\int_D \mathbf{1}_{A}(x)\mathbf{1}_{A^c}(y)\eta_\varepsilon\left(x-y\right)\rho(x)\rho(y)dydx\\
&\sim C\int_{D\cap\partial A}\rho^2(x)\,\mathrm{d}S(x).
\end{aligned}$$

The factor $1/(n^2\varepsilon^d)$ in front of the cut above is accounted for in the way we scale the cuts, see (5.6). We remark that the above just provides a rough heuristic idea as to what weight should be expected. It does not serve as a basis for our proof, since the optimal balanced graph cuts $\{Y_n, Y_n^c\}$ (minimizer of (1.1)) could be rather different from $\{A \cap X_n, A^c \cap X_n\}$ where $\{A, A^c\}$ is the optimal balanced domain cut (minimizer of (1.5)).

The reason for the presence of ρ in (1.8) is clear since the particles are drawn from the measure, ν, with density ρ, and thus the empirical measures of the sample, ν_n, converge to ν. Let us now indicate the reason for the presence of ρ^2 in (1.10). Namely for the graph weights given by (1.12) and A, X_n, and A_n as above

$$\begin{aligned}\frac{1}{n^2\varepsilon^d}\deg(A_n) &= \frac{1}{n^2}\sum_{\mathbf{x}_i\in A_n}\sum_{\mathbf{x}_j\in X_n}\frac{1}{\varepsilon^d}\eta\left(\frac{\mathbf{x}_i-\mathbf{x}_j}{\varepsilon}\right) = \int_D\int_D \mathbf{1}_{A_n}(x)\eta_\varepsilon(x-y)\,d\nu_n(x)d\nu_n(y)\\ &\sim \int_D\int_D \mathbf{1}_A(x)\eta_\varepsilon(x-y)\,\rho(x)\rho(y)dydx\\ &\sim C_\eta\int_D \mathbf{1}_A(x)\rho^2(x)dx.\end{aligned}$$

Therefore,

$$\frac{\deg(A_n)}{\deg(X_n)} \sim \frac{1}{\int_D \rho^2(x)dx}\int_D \mathbf{1}_A(x)\rho^2(x)dx = |A|_{\rho^2}.$$

Since the proofs are analogous in most of the paper, we only consider the ratio and Cheeger cuts in detail and only comment briefly on sparsest and normalized cuts.

Remark 1 (Optimality of scaling of ε_n for $d \geq 3$) *If $d \geq 3$ then the rate presented in (1.13) is sharp in terms of scaling. Namely for $D = (0,1)^d$, and ν the Lebesgue measure on D and η compactly supported, it is known from graph theory (see (Goel et al., 2004; Gupta and Kumar, 1999; Penrose, 1999)) 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 remark that the convergence fails because the lack of connectedness of random geometric graphs (with connectivity radius below the before mentioned threshold) leads to undesirable partitions. Considering different objective functionals which are still based on perimeter, but more strongly penalize existence of small connected components, or considering different graph constructions (for example by restricting attention to the giant component) could lead to convergence even for some scaling ε_n below the connectivity threshold $\frac{1}{n^{1/d}} \ll \varepsilon_n < c\frac{(\log n)^{1/d}}{n^{1/d}}$.

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.

1.4 Outline

In Section 2 we introduce the notion of convergence we use to bridge between discrete and continuum partitions. In particular this notion of convergence allows us to consider the discrete and continuum objective functionals in a common metric space, which we denote by TL^1. This notion of convergence relies on some of the notions of the theory of optimal transportation which we recall. We also recall results on optimal min-max matching between the random sample and the underlying measure (Proposition 5), which are needed in the proof of the convergence. They represent the main estimates which account for randomness. The rest of the arguments in the paper are not probabilistic.

In Section 3 we study more carefully the notion of continuum partitioning (1.5). We introduce the notion of total variation of functions on D in Subsection 3.1 and recall some of its basic properties. This enables us to introduce, in Subsection 3.2, the general setting for problem (1.5) 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 (Theorem 9) and the multi-way cuts (Theorem 12). Proving that minimizers of discrete balanced cuts converge to optimal continuum balanced cuts is reduced to proving that the discrete balanced-cut objective functionals converge (in the sense of the notion of variational convergence known as Γ-convergence) to continuum balanced-cut objective functionals. 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 results. Section 6 contains the proof of the Theorem 9 and Section 7 the proof of Theorem 12. Finally, in Section 8 we present numerical experiments which illustrate our results; we also investigate the issues related to Remark 2.

2. From Discrete to Continuum

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots$ be a sequence of i.i.d random points drawn from an underlying ground-truth measure ν. For the two-class case, our main result shows that a sequence of partitions $\{Y_n, Y_n^c\}$ of the point clouds $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$ 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 $\{Y_n, Y_n^c\}$ are the characteristic functions of Y_n and Y_n^c, namely $\mathbf{1}_{Y_n} : X_n \to \{0,1\}$ and $\mathbf{1}_{Y_n^c} : X_n \to \{0,1\}$. 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 $A^c = D \backslash A$ can be described via the characteristic functions $\mathbf{1}_A : D \to \{0,1\}$ and $\mathbf{1}_{A^c} : D \to \{0,1\}$. These too can be considered as L^1 functions, but with respect to the measure ν rather than ν_n.

We compare the partitions $\{Y_n, Y_n^c\}$ 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 García Trillos and Slepčev (2016). 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

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

This is the distance that we use to compare L^1 functions with respect to different measures.

It is motivated by optimal transportation distances (such as the Wasserstein distance and the earth-mover distance, see (García Trillos and Slepčev, 2016, 2015; Villani, 2003) and references therein). Indeed, the distance (2.1) can be seen as an optimal transportation distance between measures supported on the graphs of the functions f and g, as discussed in García Trillos and Slepčev (2016). To better understand it here, 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 $T : D \to D$ and $\mu \in \mathcal{P}(D)$, the *push-forward* of μ by T, denoted by $T_\sharp \mu \in \mathcal{P}(D)$ is given by:

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

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

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

$$\int_D f(y) d\theta(y) = \int_D f(T(x)) d\mu(x). \tag{2.2}$$

In order to give the desired interpretation of convergence in TL^1 we also need the notion of a stagnating sequence of transportation maps. Given $\mu_n \in \mathcal{P}(D)$, for $n = 1, \ldots$

and $\mu \in \mathcal{P}(D)$, a sequence $\{T_n\}_{n\in\mathbb{N}}$ of transportation maps between μ and μ_n (meaning that $T_{n\sharp}\mu = \mu_n$) is *stagnating* if

$$\int_D |x - T_n(x)| \mathrm{d}\mu(x) \to 0 \qquad \text{as } n \to \infty. \tag{2.3}$$

This notion is relevant to our considerations because for the measure ν and its empirical measures ν_n there exists (with probability one) a sequence of stagnating transportation maps $T_{n\sharp}\nu = \nu_n$. The idea is that as $n \to \infty$ 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 5

We now provide the desired interpretation of the convergence in TL^1, which is a part of Proposition 3.12 of García Trillos and Slepčev (2016).

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* $n \to \infty$.

(ii) $\mu_n \overset{w}{\rightharpoonup} \mu$ *and there exists a stagnating sequence of transportation maps* $T_{n\sharp}\mu = \mu_n$ *such that:*

$$\int_D |f(x) - f_n(T_n(x))| \, d\mu(x) \to 0, \text{ as } n \to \infty. \tag{2.4}$$

(iii) $\mu_n \overset{w}{\rightharpoonup} \mu$ *and for any stagnating sequence of transportation maps* $T_{n\sharp}\mu = \mu_n$ *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 $T_{n\sharp}\mu = \mu_n$ and then show the L^1 convergence of $f_n \circ T_n$ to f in $L^1(\mu)$. 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 $L^1(\mu)$.

Remark 4 *Suppose that the sequence of probability measures $\{\mu_n\}_{n\in\mathbb{N}}$ is such that $\mu_n \overset{w}{\rightharpoonup} \mu$. Let $f_n \in L^1(\mu_n)$ and let $f \in L^1(\mu)$. With a slight abuse of notation we say that $f_n \xrightarrow{TL^1} f$ whenever $(\mu_n, f_n) \xrightarrow{TL^1} (\mu, f)$. In particular when we write $f_n \xrightarrow{TL^1} f$ it should be clear what the corresponding measures μ_n, μ are.*

To obtain the scaling of (1.13) we need a stagnating sequence of transportation maps between ν and $\{\nu_n\}_{n\in\mathbb{N}}$ with precise information on the rate at which convergence (2.3) occurs. More precisely for some of our considerations we need the control of $T_n(x) - x$ in the stronger $L^\infty(\nu)$-norm, rather than in the $L^1(\nu)$-norm. Since the typical distance between nearby points is of order $n^{-1/d}$ the typical transportation distance, $T_n(x) - x$, must be at least of that order. The optimal upper bound on the $L^\infty(\nu)$-norm of $T_n - I$ however has an extra logarithmic correction. In particular in García Trillos and Slepčev (2015) it was shown that:

Proposition 5 *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 $C > 0$ (that depends on D and ρ) such that with probability one there exists a sequence of transportation maps $\{T_n\}_{n\in\mathbb{N}}$ from ν to ν_n ($T_{n\sharp}\nu = \nu_n$) and such that:*

$$\limsup_{n\to\infty} \frac{n^{1/d}\|\operatorname{Id} - T_n\|_{L^\infty(\nu)}}{(\log n)^{p_d}} \leq C, \tag{2.5}$$

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

The optimality of the upper bound is discussed in García Trillos and Slepčev (2015). If $d \geq 3$ it follows from the fact that for n large, with large probability there exists a ball of radius comparable to $((\ln n)/n)^{1/d}$ which contains none of the points $\mathbf{x}_1, \dots, \mathbf{x}_n$.

Having defined the TL^1-convergence for functions, we turn to the TL^1-convergence for partitions. When defining 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 of $\{1, \dots, R\}$. Having the previous observation in mind, the convergence of partitions is defined in a natural way.

Definition 6 *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 $r \in \{1, \dots, R\}$,*

$$\left(\nu_n, \mathbf{1}_{Y^n_{P_n(r)}}\right) \xrightarrow{TL^1} (\nu, \mathbf{1}_{A_r}) \qquad \text{as } n \to \infty.$$

We note that the definition above is equivalent to

$$\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{Y^n_{P_n(r)}}(x_i)\delta_{x_i} \xrightarrow{w} \mathbf{1}_{A_r}\,\nu \tag{2.6}$$

for all $r = 1, \dots, R$ which is analogous to the definition in (1.14) 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 $\Gamma_f = (\operatorname{Id} \times f)_\sharp \mu$ and $\Gamma_g = (\operatorname{Id} \times g)_\sharp \theta$ be the measures representing the graphs. Consider $d(\Gamma_f, \Gamma_g) := d_{TL^1}((\mu, f), (\theta, g))$. Proposition 3.3 in García Trillos and Slepčev (2016) (also see the paragraph right after Remark 3.1) implies that this distance metrizes the weak convergence of measures on the family of graph measures. Therefore the convergence of partitions of Definition 6 is equivalent to one given in (2.6).

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 minimizes is a minimizer of the continuum balanced cut energy. In Proposition 21 and Lemma 23 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 $f|_{\mathrm{supp}(\mu_n)}$ 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 $\rho : D \to (0, \infty)$ 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 $\Lambda \geq \lambda > 0$.

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

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

where in the above $C_c^1(D : \mathbb{R}^d)$ denotes the set of C^1-functions from D to $\mathbb{R}^d$, whose support is compactly contained in D. If u is regular enough then the weighted total variation can be written as

$$TV(u;\nu) = \int_D |\nabla u| \rho^2(x) \, \mathrm{d}x. \tag{3.2}$$

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

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

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 $TV(\cdot;\nu)$ 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 $TV(u;\nu) < \infty$, then u is a BV function and hence it has a distributional derivative Du which is a Radon measure (see Chapter 13 of Leoni (2009)). We denote by $|Du|$ the total variation of the measure Du and denote by $|Du|_{\rho^2}$ the measure determined by

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

By Theorem 4.1 of Baldi (2001)

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

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

$$TV(u;\nu) \leq \liminf_{k\to\infty} TV(u_k;\nu). \tag{3.6}$$

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

$$TV(u;\nu) = \int_{\mathbb{R}} TV(\mathbf{1}_{\{u>t\}};\nu)\ \mathrm{d}t,$$

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 Bellettini, Bouchitté, and Fragalà (1999). For a proof of the formula in the case that ρ is constant see Leoni (2009).

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

3.2 Continuum partitioning

We use the total variation to rigorously formulate the continuum partitioning problem (1.5). The precise definition of the $\mathrm{Cut}_\rho(A, A^c)$ functional in (1.5) is

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

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

Recall that $|D| = \nu(D)$, as defined in (1.8). Given that ν is a probability measure supported on D we have $|D| = 1$. We now formulate the balance terms defined by (1.7) and (1.8) using characteristic functions. We start by extending the balance term to arbitrary functions $u \in L^1(\nu)$:

$$B_{\mathrm{R}}(u) = \int_D |u(x) - \mathrm{mean}_\rho(u)|\rho(x)\ \mathrm{d}x \quad \text{and} \quad B_{\mathrm{C}}(u) = \min_{c\in\mathbb{R}} \int_D |u(x) - c|\rho(x)\ \mathrm{d}x, \tag{3.7}$$

where $\text{mean}_\rho(u)$ denotes the mean/expectation of $u(x)$ with respect to the measure $\mathrm{d}\nu = \rho \mathrm{d}x$.

Analogously, using the symbol $⨍_D f(x)\rho^2(x)dx := \frac{1}{\int_D \rho^2(x)dx}\int_D f(x)\rho^2(x)dx$, we define

$$B_{\mathrm{S}}(u) = ⨍_D |u(x) - \text{mean}_{\rho^2}(u)|\rho^2(x)\,\mathrm{d}x \quad \text{and} \quad B_{\mathrm{N}}(u) = \min_{c\in\mathbb{R}} ⨍_D |u(x)-c|\rho^2(x)\,\mathrm{d}x, \tag{3.8}$$

where $\text{mean}^2_\rho(u) = ⨍_D u(x)\rho^2(x)dx$.

We have the desired relation with balance terms defined in (1.2) and (1.3)

$$B_{\mathrm{I}}(\mathbf{1}_A) = \text{Bal}_{\mathrm{I}}(A, A^c) \text{ for } I = R, C, S, \text{ and } N \tag{3.9}$$

for every measurable subset A of D. From here on, we use B to represent B_{R}, B_{C}, B_{S}, or B_{N}, depending on the context. 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

$$\text{Ind}(D) := \left\{u \in L^1(\nu) : u = \tilde{\mathbf{1}}_A \text{ for some measurable set } A \subseteq D \text{ with } B(\mathbf{1}_A) \neq 0\right\}. \tag{3.10}$$

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

$$TV(u) = TV(\tilde{\mathbf{1}}_A) = TV\left(\frac{\mathbf{1}_A}{B(\mathbf{1}_A)}\right) = \frac{TV(\mathbf{1}_A)}{B(\mathbf{1}_A)} = \frac{\text{Cut}_\rho(A, A^c)}{\text{Bal}(A, A^c)}. \tag{3.11}$$

Consequently the problem (1.5) is equivalent to minimizing $E : TL^1(D) \to (-\infty, \infty]$, given by

$$E(\mu, u) := \begin{cases} TV(u) & \text{if } \mu = \nu \text{ and } u \in \text{Ind}(D) \\ +\infty & \text{otherwise.} \end{cases} \tag{3.12}$$

where μ is a probability measure on D, $u \in L^1(\mu)$, $TV(u) = TV(u;\nu)$, is given by (3.5) and $\text{Ind}(D)$ is defined by (3.10). Since the functional E is only non-trivial when $\mu = \nu$, from now on we write $E(u)$ instead of $E(\nu, u)$.

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

Lemma 7 *(i) The balance functions B_I are continuous on $L^1(\nu)$.*

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

Proof Let us start by proving (i). We first consider the balance term $B_{\mathrm{C}}(u)$ that corresponds to the Cheeger cut. Let $u_1, u_2 \in L^1(\nu)$. Let c_i be the median of u_i for $i = 1, 2$, that is let c_i be a minimizer of $c \mapsto \int_D |u_i(x) - c|\ \rho(x)\,\mathrm{d}x$. Then, by (3.7),

$$\begin{aligned} B(u_1) - B(u_2) &\leq \int |u_1 - c_2|\rho(x)\,\mathrm{d}x - \int |u_2 - c_2|\ \rho(x)\,\mathrm{d}x \\ &\leq \int |u_1 - u_2|\rho(x)\,\mathrm{d}x = \|u_1 - u_2\|_{L^1(\nu)}. \end{aligned}$$

Exchanging the role of u_1 and u_2 in this argument implies that $|B(u_1) - B(u_2)| \leq \|u_1 - u_2\|_{L_1(\nu)}$, which implies Lipschitz continuity of B_{C}.

Now consider the balance term $B_{\mathrm{R}}(u)$ that corresponds to the ratio cut. Let $\{u_k\}_{k=1,\dots}$ be a sequence in $L^1(\nu)$ converging to u. The inequality $||a| - |b|| \leq |a - b|$ implies that

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

Since $u_k \to u$ in $L^1(\nu)$ we have that $\mathrm{mean}_\rho(u_k) \to \mathrm{mean}_\rho(u)$ and therefore $|B_{\mathrm{R}}(u_k) - B_{\mathrm{R}}(u)| \leq \|u_k - u\|_{L_1(\nu)} + |\,\mathrm{mean}_\rho(u_k) - \mathrm{mean}_\rho(u)| \to 0$ as desired.

In order to prove (ii) suppose that $\{u_k\}_{n\in\mathbb{N}}$ is a sequence in $\mathrm{Ind}(D)$ converging in $L^1(\nu)$ to some $u \in L^1(\nu)$, we need to show that $u \in \mathrm{Ind}(D)$. By (i) we know that $B(u_k) \to B(u)$ as $k \to \infty$. Since $u_k \in \mathrm{Ind}(D)$, in particular $B(u_k) = 1$. Thus, $B(u) = 1$. On the other hand, $u_k \in \mathrm{Ind}(D)$ 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 $u \in \mathrm{Ind}(D)$. ■

Lemma 8 *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* (3.12).

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 $u_n = \tilde{\mathbf{1}}_{A_n} \in \mathrm{Ind}(D)$ converging in $L^1(\nu)$ to $u \in L^1(\nu)$. From Lemma 7 it follows that $u \in \mathrm{Ind(D)}$ and hence $u = \tilde{\mathbf{1}}_A$ for some A with $B(\mathbf{1}_A) > 0$. Therefore $\mathbf{1}_{A_n} \to \mathbf{1}_A$ as $n \to \infty$ 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 $B(\mathbf{1}_A) > 0$, $1/B(\mathbf{1}_{A_n}) \to 1/B(\mathbf{1}_A)$ as $n \to \infty$.

The pre-compactness of any minimizing sequence of (3.12) follows directly from Theorem 5.1 of Baldi (2001), which completes the proof. ■

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 $\rho : D \to \mathbb{R}$ be a continuous density which is bounded below and above by positive constants, that is, for all $x \in D$

$$\lambda \leq \rho(x) \leq \Lambda \tag{4.1}$$

for some $\Lambda \geq \lambda > 0$. We let ν be the measure $\mathrm{d}\nu = \rho \mathrm{d}x$. Let $\boldsymbol{\eta} : [0, \infty) \to [0, \infty)$ be the radial profile of the similarity kernel, namely the function satisfying $\eta(x) = \boldsymbol{\eta}(|x|)$. We assume

(K1) $\boldsymbol{\eta}(0) > 0$ and $\boldsymbol{\eta}$ is continuous at 0.

(K2) $\boldsymbol{\eta}$ is non-increasing.

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

We refer to the quantity $\sigma_{\boldsymbol{\eta}}$ as the *surface tension* associated to $\boldsymbol{\eta}$. In the above, $\langle x, e_1\rangle$ denotes the inner product of the vector x with the vector whose first entry is 1 and whose other entries are equal to zero. We remark that due to radial symmetry, the vector e_1 can be replaced by any unit vector in $\mathbb{R}^d$ without changing the value of $\sigma_{\boldsymbol{\eta}}$.

The kernel $\eta : \mathbb{R}^d \to [0, \infty)$ is now given by $\eta(x) = \boldsymbol{\eta}(|x|))$.

These hypotheses on $\boldsymbol{\eta}$ hold for the standard similarity functions used in clustering contexts, such as the Gaussian similarity function $\boldsymbol{\eta}(r) = \exp(-r^2)$ and the proximity similarity kernel ($\boldsymbol{\eta}(r) = 1$ if $r \leq 1$ and $\boldsymbol{\eta}(r) = 0$ otherwise). For a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from the measure ν, we denote by ν_n the empirical measure associated to the sample.

The main result of our paper is:

Theorem 9 (Consistency of cuts) *Let domain D, probability measure ν,with density ρ, and kernel η satisfy the conditions above. Let ε_n denote any sequence of positive numbers converging to zero that satisfy*

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

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

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

where $\boldsymbol{\eta}$ satisfies assumptions (K1)-(K3). Finally, let $\{Y_n^*, Y_n^{*c}\}$ denote any optimal balanced cut of $\mathcal{G}_n$ (solution of problem (1.1)). If problem (3.12) has a unique solution $\{A^*, A^{*c}\}$, 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 (3.12) then with probability one, $\{Y_n^*, Y_n^{*c}\}$ converges along a subsequence to an optimal continuum balanced cut.

Additionally, with probability one, $\mathcal{C}_n$ the minimum balanced cut of the graph $\mathcal{G}_n$ (the minimum of (1.1)), satisfies

$$\lim_{n\to\infty} \frac{2\mathcal{C}_n}{n^2 \varepsilon_n^{d+1}} = \sigma_\eta \mathcal{C}, \tag{4.2}$$

where σ_η is the surface tension associated to the kernel η and $\mathcal{C}$ is the minimum of (1.5).

As the proofs for sparsest and normalized cuts are analogous, in the remainder of the paper we only treat the ratio and Cheeger cuts in detail.

Remark 10 *For simplicity of notation, from now on we make the assumption that problem* (3.12) *has a unique solution* $\{A^*, A^{*c}\}$. *In the general case, Theorem 9 follows using the same approach; the only difference is that the convergence of minimizers happens along subsequences (see Proposition 17 below).*

As we discussed in Remark 1 for $d \geq 3$ the scaling of $\varepsilon = \varepsilon_n$ on n is essentially the best possible. The proof of Theorem 9 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 uses the results obtained of García Trillos and Slepčev (2016), where the notion of Γ-convergence was introduced in the context of objective functionals on random data samples, 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 11 *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 5. 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*

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

This important property follows from a general result on Γ*-convergence which we recall in Proposition 17.*

We also establish the following multi-class equivalent to Theorem 9.

Theorem 12 *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 9. Let* $(Y^{*n}_1, \ldots, Y^{*n}_R)$ *denote any optimal balanced cut of* $\mathcal{G}_n$, *that is a minimizer of* (1.4). *If* $(A_1^*, \ldots, A_R^*)$ *is the unique optimal balanced cut of* D *(that is minimizer of* (1.11)*) then with probability one the sequence* $(Y^{*n}_1, \ldots, Y^{*n}_R)$ *converges to* $(A_1^*, \ldots, A_R^*)$ *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* (1.4), *satisfies*

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

where σ_η *is the surface tension associated to the kernel* η *and* $\mathcal{C}$ *is the minimum of* (1.11).

The proof of Theorem 12 involves modifying the geometric measure theoretical results of García Trillos and Slepčev (2016). This leads to a substantially longer and more technical proof than the proof of Theorem 9, 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 11 apply to Theorem 12.

5. Background on Γ-convergence

We recall and discuss the notion of Γ-convergence. The usual Γ-convergence is defined for deterministic functionals. It extends to the random functionals that we consider in a natural way. Namely for almost every realization of the random event (in our case a sequence of random points in the domain) we require the Γ-convergence of resulting, deterministic, functionals. Such notion of Γ convergence has been used by Dirr and Orlandi (2009). We now define it precisely.

Let (X, d_X) be a metric space and let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space. Let $F_n : X \times \Omega \to [0, \infty]$ be a sequence of random functionals. For brevity, instead of writing $F_n(x, \omega)$ we simply write $F_n(x)$ with understanding that an element $\omega \in \Omega$ has been fixed.

Definition 13 *The sequence of random functionals $\{F_n\}_{n\in\mathbb{N}}$ Γ-converges with respect to metric d_X to the deterministic functional $F : X \to [0, \infty]$ as $n \to \infty$ if for $\mathbb{P}$-almost every ω, 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,*

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

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

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

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 14 *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 $F(x_k) \to F(x)$ as $k \to \infty$, then the limsup inequality is true for every $x \in X$. The proof of the claim is straightforward, using, for example Theorem 1.17(iii) of Braides (2002). This property is not related to the randomness of the functionals in any way.*

Definition 15 *We say that the sequence of nonnegative random functionals $\{F_n\}_{n\in\mathbb{N}}$ satisfies the compactness property if for $\mathbb{P}$-almost every ω , the following statement holds: any sequence $\{x_n\}_{n\in\mathbb{N}}$ bounded in X and for which*

$$\limsup_{n\to\infty} F_n(x_n) < +\infty,$$

is relatively compact in X.

Remark 16 *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 (Braides, 2002; Dal Maso, 1993) in the deterministic setting . We present its proof in this random setting 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 17 *Let $F_n : X \times \Omega \to [0,\infty]$ 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 $F : X \to [0,\infty]$ which is not identically equal to $+\infty$. Suppose that for $\mathbb{P}$ almost every ω there is a bounded sequence $\{x_n\}_{n\in\mathbb{N}}$ (which may depend on ω) satisfying*

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

Then, with $\mathbb{P}$-probability one,

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

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, 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 $\omega \in \Omega'$.

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 $\tilde{x}_n \to \tilde{x}$ and such that

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

By 5.1 we deduce that

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

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

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

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 $\limsup_{n\to\infty} F_n(x_n) < +\infty$. 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 $x_n \to x^*$. By the liminf inequality, we deduce that

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

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). ■

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 García Trillos and Slepčev (2016). 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 $GTV_{n,\varepsilon_n} : TL^1(D) \to [0,\infty]$ the functional defined as follows: $GTV_{n,\varepsilon_n}(\mu, u_n) = \infty$ if $\mu \neq \nu_n$ and

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

where η is a kernel satisfying conditions **(K1)-(K3)**. Since we consider GTV_{n,ε_n} only for $\mu = \nu_n$, from now on we only write $GTV_{n,\varepsilon_n}(u_n)$, instead of $GTV_{n,\varepsilon_n}(\nu_n, u_n)$. Using the empirical measure ν_n, we may alternatively write $GTV_{n,\varepsilon_n}(u_n)$ as

$$GTV_{n,\varepsilon_n}(u_n) = \frac{1}{\varepsilon_n^{d+1}} \int\int \eta\left(\frac{|x-y|}{\varepsilon_n}\right) |u_n(x) - u_n(y)| d\nu_n(x) d\nu_n(y).$$

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,

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

we recall that $w_{ij} = \eta\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\varepsilon_n}\right)$.

Now we recall the Theorems 1.1 and 1.2 of García Trillos and Slepčev (2016).

Theorem 18 (Γ - Convergence) *Let the domain D, measure ν, kernel η, sample points $\{\mathbf{x}_i\}_{i\in N}$, sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$, and graph $\mathcal{G}_n$ satisfy the assumptions of Theorem 9. Then,*

GTV_{n,ε_n}, defined by (5.6), *Γ-converge to $\sigma_\eta TV_\nu$ as $n \to \infty$ in the TL^1 sense, where σ_η is the surface tension associated to the kernel η (see condition (**K3**)) and TV_ν is the extension to $TL^1(D)$ of weighted (by ρ^2) total variation functional introduced in* (3.1), *defined as follows:*

$$TV_\nu((u,\mu)) = \begin{cases} \sigma_\eta TV(u) & \text{if } \mu = \nu \\ +\infty & \text{else.} \end{cases}$$

Moreover, we have the following compactness result.

Theorem 19 (Compactness) *Under the hypothesis of Theorem 18, the sequence of functionals $\{GTV_{n,\varepsilon_n}\}_{n\in\mathbb{N}}$ satisfies the compactness property. Namely, for $\mathbb{P}$-almost every ω the following holds: if a sequence $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in L^1(\nu_n)$ satisfies*

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

and

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

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

To conclude this section, we present Corollary 1.3 in García Trillos and Slepčev (2016), which allows us to restrict the functionals GTV_{n,ε_n} and TV to characteristic functions of sets and still obtain Γ-convergence. Observe that the only subtle point is the limsup inequality as the liminf inequality and compactness statements are particular cases of Theorem 18 and Theorem 19.

Theorem 20 *Under the assumptions of Theorem 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

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

The results stated above are the main tools in order to establish our main theorems. In the next section we use them together with a careful treatment of the balance term appearing in the denominator of the Cheeger/ratio cut functional.

6. Consistency of two-way balanced cuts

Here we prove Theorem 9.

6.1 Outline of the proof

Before proving that minimal balanced cuts $\{Y_n^*, Y_n^{*c}\}$ converge to minimal continuum partitions $\{A^*, A^{*c}\}$ in the sense of Definition 6, we first pause to outline the main ideas. Rather than work directly with the graph-cut-based functional defined on the sets of vertices we work with its relaxation defined on the set of functions from the graph to reals, $L^1(\nu_n)$. The relaxed discrete functionals E_n are defined in (6.6) and the relaxed continuum one, E is defined in (3.12).

We first show, by an explicit construction in Subsection 6.2, that the rescaled indicator functions of minimal balanced cuts, $\tilde{\mathbf{1}}_{Y_n}(x) := \alpha_n \mathbf{1}_{Y_n}(x)$, (for explicit coefficient α_n that we will define later),

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

Similarly, in Subsection 3.2 we showed that the normalized indicator functions

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

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 23 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 17). If $\{A^*, A^{*c}\}$ is the unique minimizer, the convergence of the sequences (up to relabeling) $\{u_n^*\}, \{u_n^{**}\}$ follows. The convergence of the partition $\{Y_n^*, Y_n^{*c}\}$ toward the partition $\{A^*, A^{*c}\}$ in the sense of Definition 6 is a direct consequence. The convergence (4.2) follows from (5.2) in Proposition 17.

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

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

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

$$B_{\mathrm{R}}^n(\mathbf{1}_{Y_n}) = \mathrm{Bal}_{\mathrm{R}}(Y_n, Y_n^c), \quad B_{\mathrm{C}}^n(\mathbf{1}_{Y_n}) = \mathrm{Bal}_{\mathrm{C}}(Y_n, Y_n^c). \tag{6.4}$$

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

$$\tilde{\mathbf{1}}_{Y_n}(x) = \mathbf{1}_{Y_n}(x)/B_{\mathrm{C}}^n(\mathbf{1}_{Y_n}) \quad \text{or} \quad \tilde{\mathbf{1}}_{Y_n}(x) = \mathbf{1}_{Y_n}(x)/B_{\mathrm{R}}^n(\mathbf{1}_{Y_n}).$$

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

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

Now, suppose that $u_n \in \mathrm{Ind}_n(D)$, in other words that $u_n = \tilde{\mathbf{1}}_{Y_n}$, for some set Y_n with $B_n(\mathbf{1}_{Y_n}) > 0$. Using (3.9) together with the fact that GTV_{n,ε_n} (defined in (5.6)) is one-homogeneous implies, as in (3.11)

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

Thus, minimizing GTV_{n,ε_n} over all $u_n \in \mathrm{Ind}_n(D)$ 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, a proper reformulation of (1.1) defined over $TL^1(D)$ instead of subsets of X_n. The task is to

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

Since the measure is clear from context, from now on we write $E_n(u_n)$ for $E_n(\nu_n, u_n)$.

6.3 Γ-Convergence

Proposition 21 (Γ-Convergence) *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 9. Let E_n be as defined in* (6.7) *and E as in* (3.12)*. Then*

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

*where σ_η is the surface tension defined in assumption (**K3**). This is implied by the following:*

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

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

We leverage Theorem 18 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 (3.12).

Lemma 22 *With probability one, the following hold:*

(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 $B_n(u_n) \to B(u)$.

(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 $\{\nu_n\}_{n\in\mathbb{N}}$ (one such sequence exists with probability one by Proposition 5) . Then, we have $u_n \circ T_n \xrightarrow{L^1(\nu)} u$ and thus by Lemma 7, we have that $B(u_n \circ T_n) \to B(u)$. To conclude the proof we notice that $B(u_n \circ T_n) = B_n(u_n)$ for every n. Indeed, by the change of variables (2.2) we have that for every $c \in \mathbb{R}$

$$\int_D |u_n(x) - c| \mathrm{d}\nu_n(x) = \int_D |u_n \circ T_n(x) - c| \mathrm{d}\nu(x). \tag{6.10}$$

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 $\mathrm{mean}_n(u_n) = \mathrm{mean}_\rho(u_n \circ T_n)$ 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. ■

Now we turn to the proof or Proposition 21.

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

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

First assume that $u \in \mathrm{Ind}(D)$. 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 $E_n(u_n) < +\infty$, which in particular implies that $E_n(u_n) = GTV_{n,\varepsilon_n}(u_n)$. Theorem 18 then implies that

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

Now let as assume that $u \notin \mathrm{Ind}(D)$. 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 7 , the set $\mathrm{Ind}(D)$ is a closed subset of $L^1(\nu)$. We conclude that $u_n \circ T_n \notin \mathrm{Ind}(D)$ for all large enough n. From the proof of Lemma 22 we know that $B_n(u_n) = B(u_n \circ T_n)$ and from this fact, it is straightforward to show that $u_n \circ T_n \notin \mathrm{Ind}(D)$ if and only if $u_n \notin \mathrm{Ind}_n(D)$. Hence, $u_n \notin \mathrm{Ind}_n(D)$ 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

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

Let us start by assuming that $u \notin \mathrm{Ind}(D)$. In this case $E(u) = +\infty$. From Theorem 18 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 $u \in \mathrm{Ind}(D)$, we know that $u = \tilde{\mathbf{1}}_A$ for some measurable subset A of D with $B(\mathbf{1}_A) \neq 0$. By Theorem 20, 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

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

Since $\mathbf{1}_{Y_n} \xrightarrow{TL^1} \mathbf{1}_A$ Lemma 22 implies that

$$B_n(\mathbf{1}_{Y_n}) \to B(\mathbf{1}_A). \tag{6.12}$$

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

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

Since, $u_n \in \mathrm{Ind}_n(D)$ 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 $u \in \mathrm{Ind}(D)$, we have $E(u) = TV(u)$. These facts together with the previous chain of inequalities imply the result. ■

6.4 Compactness

Lemma 23 (Compactness) *With probability one the following statement holds: Any sequence $\{u_n\}_{n\in\mathbb{N}}$ with*

$$\limsup_{n\to\infty} E_n(u_n) < +\infty$$

is precompact in TL^1. In particular, any sequence $\{u_n^*\}_{n\geq 1}$, of minimizers of E_n (defined in (6.1) and (6.2)) are precompact in the TL^1-sense.

Proof Let u_n denote a sequence satisfying

$$\limsup_{n\to\infty} E_n(u_n) < \infty.$$

To show that any subsequence of u_n has a convergent subsequence it suffices to show that both

$$\limsup_{n\to\infty} \; GTV_{n,\varepsilon_n}(u_n) < +\infty \tag{6.13}$$

$$\limsup_{n\to\infty} \; \|u_n\|_{L^1(\nu_n)} < +\infty \tag{6.14}$$

hold due to Theorem 19. Since the result is about asymptotic behavior, we can assume without loss of generality that $\sup_{n\in\mathbb{N}} E_n(u_n) < +\infty$. Inequality (6.13) follows from the fact that $E_n(u_n) = GTV_{n,\varepsilon_n}(u_n)$. Note that $E_n(u_n) < \infty$ in particular implies that u_n has the form $u_n = \frac{\mathbf{1}_{Y_n}}{B_n(Y_n)}$ for some $Y_n \subseteq X_n$.

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 $v_n = \frac{\mathbf{1}_{Y_n^c}}{B_n(Y_n^c)}$ 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. Thus, it follows from Proposition 21 , that

$$\sigma_\eta E(v) \leq \liminf_{k\to\infty} E_{n_k}(v_{n_k}) < \infty,$$

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

$$\frac{1}{B_{n_k}(Y_{n_k})} \to \frac{1}{B(A)}.$$

Therefore, for large enough k we have

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

We conclude that $\|u^{n_k}\|_{L^1(\nu_{n_k})}$ remains bounded in L^1, so that it satisfies (6.14) and (6.13) simultaneously. This yields compactness in the Cheeger cut case.

Now consider the balance term $B(u) = B_{\mathrm{R}}(u)$ that corresponds to the ratio cut. Define a sequence $v_n := u_n - \mathrm{mean}_n(u_n)$, 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

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

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 $v_{n_k} \circ T_{n_k}(x) \to v(x)$ for ν-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

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

and also

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

In particular the ν-measure of the region in which v is positive is strictly greater than zero, and likewise the ν-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

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

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

6.5 Conclusion of the proof of Theorem 9

We may now turn to the final step of the proof. From Proposition 17, 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 22, while $\mathbf{1}_{Y_{n_k}^*} \xrightarrow{TL^1} \mathbf{1}_{A^{*c}}$ 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 $\mathbf{1}_{Y_{n_k}^c} \xrightarrow{TL^1} \mathbf{1}_{A^*}$. Thus in either case we have

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

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 12.

Just as what we did in the two-class case, the first step in the proof of Theorem 12 involves a reformulation of both the balanced graph-cut problem (1.4) and the analogous balanced domain-cut problem (1.11) 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 (3.10) respectively.

We can then let the sets $\mathcal{M}_n(D)$ and $\mathcal{M}(D)$ consist of those collections $\mathcal{U} = (u_1, \ldots, u_R)$ 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

$$\mathcal{M}_n(D) = \left\{ (u_1^n, \ldots, u_R^n) \ : u_r^n \in \mathrm{Ind}_n(D), \int_D u_r^n(x) u_s^n(x) \, \mathrm{d}\nu_n(x) = 0 \ \text{ if } \ r \neq s, \ \sum_{r=1}^R u_r^n > 0 \right\} \tag{7.1}$$

$$\mathcal{M}(D) = \left\{ (u_1, \ldots, u_R) \ : u_r \in \mathrm{Ind}(D), \int_D u_r(x) u_s(x) \, \mathrm{d}\nu(x) = 0 \ \text{ if } \ r \neq s, \ \sum_{r=1}^R u_r > 0 \right\}. \tag{7.2}$$

Note for example that if $\mathcal{U} = (u_1, \ldots, u_R) \in \mathcal{M}(D)$, 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 $\{A_1, \ldots, A_R\}$ 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 $\{A_1, \ldots, A_R\}$ cover D up to Lebesgue-null sets.

We proceed to define the functionals on the space of R-tuples of L^1 functions, namely

$$TL^1(D, R) := \left\{ (\mu, \mathcal{U}) \ : \ \mu \in \mathcal{P}(D), \mathcal{U} = (u_1, \ldots, u_R), \, u_i \in L^1(\mu) \text{ for } i = 1, \ldots, R \right\}.$$

We note that convergence in $TL^1(D, R)$ is equivalent to convergence of the R components in $TL^1(D)$.

One may follow the same argument in the two-class case to conclude that the minimization

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

is equivalent to the balanced graph-cut problem (1.4), while the minimization

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

is equivalent to the balance domain-cut problem (1.11).

As in the two-class case we omit the first argument of E_n and E, when it is clear from context.

At this stage, the proof of Theorem 12, 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).

Proposition 24 (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 9. Consider functionals E_n of* (7.3) *and E of* (7.4)*. Then*

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

That is, with probability one, all of the following statements hold

1. *For any* $\mathcal{U} \in [L^1(\nu)]^R$ *and any sequence* $\mathcal{U}_n \in (L^1(\nu_n))^R$ *that converges to* $\mathcal{U}$ *in the* TL^1 *sense,*
$$E(\mathcal{U}) \leq \liminf_{n\to\infty} E_n(\mathcal{U}_n). \tag{7.5}$$

2. *For any* $\mathcal{U} \in [L^1(\nu)]^R$ *there exists a sequence* $\mathcal{U}_n \in (L^1(\nu_n))^R$ *that both, converges to* $\mathcal{U}$ *in the* TL^1*-sense, and also satisfies*
$$\limsup_{n\to\infty} E_n(\mathcal{U}_n) \leq E(\mathcal{U}). \tag{7.6}$$

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 Lemma 23.

Lemma 25 (Compactness) *With probability one the following statement holds: Any sequence* $\{\mathcal{U}_n\}_{n\in\mathbb{N}}$ *with* $\mathcal{U}_n \in [L^1(\nu_n)]^R$ *satisfying*
$$\limsup_{n\to\infty} E_n(\mathcal{U}_n) < +\infty,$$
is precompact in the TL^1*-sense. In particular, 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 24 and Lemma 25, the arguments presented in Subsections 6.1 and 6.5 can be adapted in a straightforward way to complete the proof of Theorem 12. So we focus on the proof of Proposition 24, 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 24

The next lemma is the multiclass analogue of Lemmas 7 and 22 combined.

Lemma 26 (i) *If* $\mathcal{U}_k \to \mathcal{U}$ *in* $(L^1(\nu))^R$ *then* $B(u_r^k) \to B(u_r)$ *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* $B_n(u_r^n) \to B(u_r)$ *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) follow directly from the proof of Proposition 22. Statement (iv) is exactly as in Proposition 22.

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 $u_r^k \in \mathrm{Ind}(D)$ for every $k \in \mathbb{N}$, and since $\mathrm{Ind}(D)$ is a closed subset of $L^1(\nu)$ (by Proposition 22), we deduce that $u_r \in \mathrm{Ind}(D)$ 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, $u_r^k \to u_r$ for almost every x in D. Hence, for $r \neq s$ we have

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

Now let us write $u_r^k = \mathbf{1}_{A_r^k}/B(\mathbf{1}_{A_r^k})$ and $u_r = \mathbf{1}_{A_r^k}/B(\mathbf{1}_{A_r})$. As in the proof of Proposition 22 we must have $B(\mathbf{1}_{A_r^k}) \to B(\mathbf{1}_{A_r})$ as $k \to \infty$. Thus, for almost every $x \in D$

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

■

Proof [of Proposition 24]

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 make the limit inferior less than infinity. In this case we easily have

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

The last inequality follows from Theorem 18. This establishes the first statement in Proposition 24.

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. It proves useful to 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.

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 14). We furthermore remark that it is enough to consider $\mathcal{U} = (u_1, \ldots, u_R) \in (L^1(D))^R$ 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 that for every $s = 1, \dots, R$,

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

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.

We say that a subset of $\mathbb{R}^d$ has a *piecewise (PW) smooth* boundary if the boundary is a subset of the union of finitely many open $d-1$-dimensional manifolds embedded in $\mathbb{R}^d$. We first construct a recovery sequence for $\mathcal{U}$, as above, 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 $|D\mathbf{1}_{B_r}|_{\rho^2}(\partial D) = 0$. 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 sets with piecewise smooth boundary. 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 5. We let A_n^r be the set for which $\mathbf{1}_{A_r^n} = \mathbf{1}_{Y_n^r} \circ T_n$.

We first notice that the fact that B_r has a piecewise smooth boundary in $\mathbb{R}^d$ and the fact that $\mathbf{1}_{A_r^n} - \mathbf{1}_{A_r}$ is nontrivial only within the tubular neighborhood of ∂B_r of radius $||\operatorname{Id} - T_n||_\infty$, imply that

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

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 Weyl (1939), page 461). In particular, note that by the change of variables (2.2) we have, $|Y_r^n| = |A_r^n| \to |A_r|$ as $n \to \infty$, 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 $u_r^n \in \operatorname{Ind}_n(D)$ 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 García Trillos and Slepčev (2016). In particular, we can assume that η has the form $\boldsymbol{\eta}(|z|) = a$ for $|z| < b$ and $\eta(z) = 0$ otherwise; the general case follows in a straightforward way by using an approximating procedure with

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

kernels that are a finite sum of step functions like the one considered previously (see the proof of Theorem 1.1 in García Trillos and Slepčev (2016)) .

We set $\tilde{\varepsilon}_n := \varepsilon_n + \frac{2}{b}||\operatorname{Id} - T_n||_\infty$. Recall that by assumption $||\operatorname{Id} - T_n||_\infty \ll \varepsilon_n$ (see the statement of Theorem 9 and Proposition 5), and thus $\tilde{\varepsilon}_n$ is a small 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

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

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 an 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

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

Since $\frac{\varepsilon_n}{\tilde{\varepsilon}_n} \to 1$, the previous inequalities imply that

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

Finally, from remark 4.3 in García Trillos and Slepčev (2016) we deduce that

$$\limsup_{n\to\infty} \widetilde{TV}_{\tilde{\varepsilon}_n}(\mathbf{1}_{A_r^n}) \leq \sigma_\eta TV(\mathbf{1}_{A_r}),$$

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

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

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

Density. To prove Proposition 24, we show that for any $\mathcal{U} = (\tilde{\mathbf{1}}_{A_1}, \ldots, \tilde{\mathbf{1}}_{A_R})$ where each of the sets A_r has finite perimeter, there exists a sequence $\left\{\mathcal{U}_m = (\tilde{\mathbf{1}}_{A_1^m}, \ldots, \tilde{\mathbf{1}}_{A_R^m})\right\}_{m\in\mathbb{N}}$, where each of the $\mathcal{U}_m$ is induced by piecewise smooth sets, and such that for every $r \in \{1, \ldots, R\}$

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

and

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

Note that in fact, by establishing the existence of such approximating sequence, it immediately follows that $\mathcal{U}_m \to \mathcal{U}$ in $(L^1(\nu))^R$ and that $\lim_{m\to\infty} E(\mathcal{U}_m) = E(\mathcal{U})$ (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 27 *Let $\{A_1, \dots, A_R\}$ denote a collection of open and bounded sets with smooth boundary in $\mathbb{R}^d$ that satisfy*

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

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

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

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 open, bounded sets with smooth boundary satisfying (7.8) . By the induction hypothesis it is enough to show that we can find $r \in \{1, \dots, R\}$ such that

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

To simplify notation, denote by Γ_i the set ∂A_i and define a_{ij} as the quantity

$$a_{ij} := \int_{\Gamma_i \cap (A_j \setminus \bigcup_{k\neq i, k\neq j} A_k) \cap D} \rho^2(x)\, \mathrm{d}\mathcal{H}^{d-1}(x).$$

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

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

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

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

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

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

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. ■

Lemma 28 *Let D denote an open, bounded domain in $\mathbb{R}^d$ with Lipschitz boundary and let $(B_1, \ldots, B_R)$ denote a collection of R bounded and mutually disjoint subsets of $\mathbb{R}^d$ that satisfy*

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

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

$$\mathbf{1}_{A_r^m} \overset{L^1(\mathbb{R}^d)}{\longrightarrow} \mathbf{1}_{B_r} \quad \textit{and} \quad \lim_{m\to\infty} TV(\mathbf{1}_{A_r^m}; \nu) = TV(\mathbf{1}_{B_r}; \nu) \tag{7.12}$$

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

Proof The proof of this lemma follows very similar ideas to those used when proving that sets with smooth boundary approximate sets with finite perimeter (see Theorem 13.46 in Leoni (2009)). Since our goal is to approximate partitions of more than two sets, we need to modify the arguments slightly. We highlight the important steps in the proof and refer to Leoni (2009) and Ambrosio et al. (2000) for details.

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 $\rho^2 : \mathbb{R}^d \to (0, \infty)$ is lower semi-continuous and bounded below and above by positive constants then, it belongs to the class of weights considered in Baldi (2001) where the weighted total variation is studied.

For $r = 1, \ldots, R$, we consider sequences of functions $u_r^k \in C^\infty(\mathbb{R}^d, [0, 1])$ satisfying

$$u_r^k \overset{L^1(\mathbb{R}^d)}{\longrightarrow} \mathbf{1}_{B_r} \text{ and } TV(u_r^k; \nu) \to TV(\mathbf{1}_{B_r}; \nu), \quad \text{as } k \to \infty. \tag{7.13}$$

This can be achieved by using standard, radially symmetric mollifiers J_k and setting $u_k^r = J_k * \mathbf{1}_{B_r}$, where $*$ stands for convolution. The functions J_k have the form $J_k(x) = k^d J(k|x|)$, where $J : [0, \infty) \to [0, \infty)$ is a smooth, decreasing function satisfying $\int_{\mathbb{R}^d} J(|x|)dx = 1$. See Theorem 13.46 in Leoni (2009) for more details.

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

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

To see this, note that the fact that D is an open and bounded set with Lipschitz boundary implies that (see Grisvard (1985), Theorem 1.2.2.2) there exists a cone $C \subseteq \mathbb{R}^d$ with non-empty interior and vertex at the origin, a family of rotations $\{R_x\}_{x\in D}$ and a number $\zeta > 0$ such that for every $x \in D$,

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

The isotropy of J_k implies that

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

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

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

for all $x \in D$ as claimed.

Now, for $k \in \mathbb{N}$, $t \in (0,1)$ and $r = 1, \ldots, R$, we let $B_r^k(t) := \{x : u_r^k(x) > t\}$. From (7.13), Sard's lemma (see Corollary 13.45 of Leoni (2009)), the lower semi-continuity of the total variation and the coarea formula for total variation, it follows that for almost every $t \in (0,1)$,

$$\partial B_r^k(t) \text{ is smooth } \forall k, \quad \lim_{k\to\infty} TV(\mathbf{1}_{B_r^k(t)};\nu) = TV(\mathbf{1}_{B_r};\nu), \quad \mathbf{1}_{B_r^k(t)} \overset{L^1(\nu)}{\longrightarrow} \mathbf{1}_{B_r}, \tag{7.14}$$

for all $r = 1, \ldots, R$.

Combining (7.14) with Lemma 2.95 in Ambrosio et al. (2000), we can find positive numbers $t_1, \ldots, t_R$ strictly smaller than α/R, such that for every $r = 1, \ldots, R$

$$\partial B_r^k(t_r) \text{ is smooth } \forall k, \quad \lim_{k\to\infty} TV(\mathbf{1}_{B_r^k(t_r)};\nu) = TV(\mathbf{1}_{B_r};\nu), \quad \mathbf{1}_{B_r^k(t_r)} \overset{L^1(\nu)}{\longrightarrow} \mathbf{1}_{B_r}, \tag{7.15}$$

and such that for $r \neq s$,

$$\mathcal{H}^{d-1}\left(\partial B_r^k(t_r) \cap \partial B_s^k(t_s)\right) = 0, \quad \forall k \in \mathbb{N}.$$

We let $B_r^k := B_r^k(t_r)$ for $r = 1, \ldots, R$ and $k \in \mathbb{N}$. We claim that for every $k \in \mathbb{N}$, the sets $B_1^k, \ldots, B_R^k$ cover D. To see this, suppose there exists $x \in D \setminus \left(\bigcup_{r=1}^R B_r^k\right)$. This would imply that $u_r^k(x) \leq t_r$ for all r. In turn, $\Sigma^k(x) \leq \sum_{r=1}^R t_r < \alpha$, which contradicts the estimate on Σ^k obtained earlier.

For every $k \in \mathbb{N}$, we can now use the sets $(B_1^k, \ldots, B_R^k)$ as input in Lemma (27) to obtain a partition $(A_1^k, \ldots, A_R^k)$ of D, defined by

$$A_r^k := B_r^k \setminus \bigcup_{s=\pi_k^{-1}(r)+1}^{R} B_{\pi_k(s)}^k$$

where π_k is a permutation of $\{1, \ldots, R\}$ guaranteeing that for every $r = 1, \ldots, R$

$$TV\left(\mathbf{1}_{A_r^k};\nu\right) \leq TV(\mathbf{1}_{B_r^k};\nu). \tag{7.16}$$

Each A_r^k has a piecewise smooth boundary due to the fact that each B_r^k has a smooth boundary. The disjointness of $(B_1, \ldots, B_R)$ combines with the L^1-convergence of $\mathbf{1}_{B_r^k}$ to $\mathbf{1}_{B_r}$ to show that $\mathbf{1}_{A_r^m} \overset{L^1(\mathbb{R}^d)}{\longrightarrow} \mathbf{1}_{B_r}$ as well. Finally, the lower semi-continuity of the total variation together with (7.16) and (7.15) imply (7.12). ∎

To complete the construction, and therefore to conclude the proof of Lemma 24, we need to verify the hypotheses (i) – (ii) of the previous lemma. This is the content of our final lemma.

Lemma 29 *Let D be an open bounded domain with Lipschitz boundary and let $\{A_1, \dots, A_R\}$ denote a disjoint collection of sets that satisfy*

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

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

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

The proof follows from Remark 3.43 in Ambrosio et al. (2000) (which with minimal modifications applies to total variation with weight ρ^2). ■

8. Numerical Experiments

We now present numerical experiments to provide a concrete demonstration and visualization of the theoretical results developed in this paper. We conduct all of our experiments using the Cheeger cut algorithm of Bresson et al. (2012); we omit the ratio cut for the sake of brevity and to avoid redundancy. These experiments focus on elucidating when and how minimizers of the graph-based Cheeger cut problem,

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

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

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

We always take $\rho(x) := 1/\mathrm{vol}(D)$ 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

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

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

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

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

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

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, converging to the optimal continuum Cheeger cut.

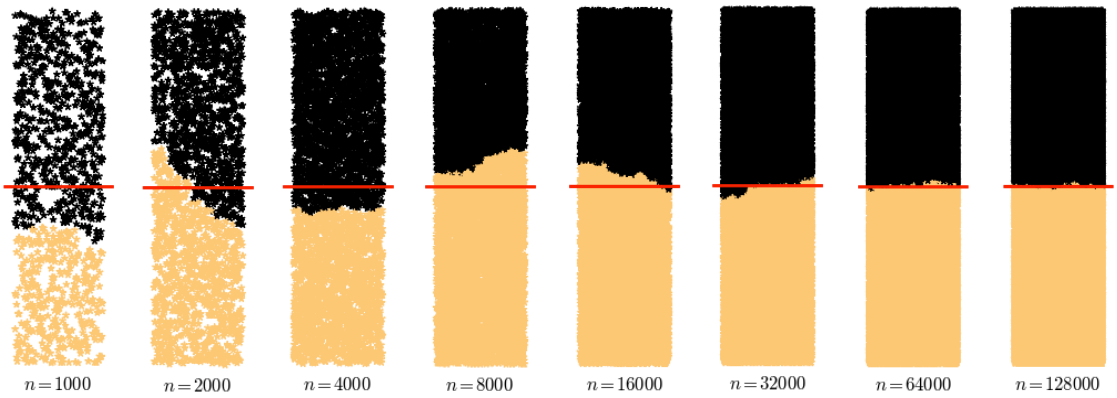


Figure 2: Visualization of the convergence process. Each figure depicts a computed optimal partition Y_n^* (in black) of one random realization of the random geometric graph $\mathcal{G}_n = (X_n, W_n)$ for each $k \in \{0, 1, \ldots, 7\}$, where $n = 1000 \times 2^k$, $\varepsilon = n^{-0.3}$ and the 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, that is the boundary of the set $A_1 := \{(x, y) \in D_1 : y > 2\}$, at the continuum level.

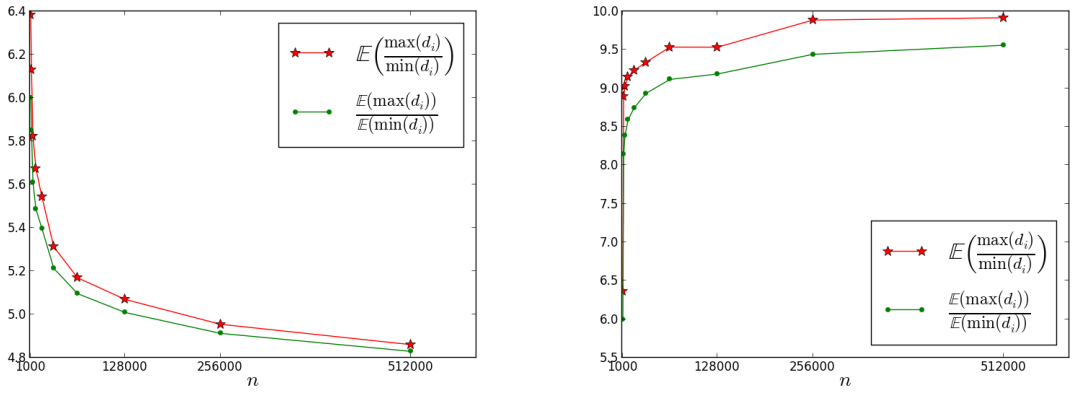
Each of our experiments use the kernel $\boldsymbol{\eta}(z) = \mathbf{1}_{\{|z| \leq 1\}}$ for the computation of the similarity weights,

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

so that the graphs $\mathcal{G}_n = (X_n, W_n)$ correspond to random geometric graphs (seePenrose (2003)). We use the domain D_1 only for the illustrations in Figure 2; all other experiments are conducted on the domain D_2. We use the steepest descent algorithm of Bresson et al. (2012) to solve the graph-based Cheeger cut problem on these graphs. This algorithm relies upon a non-convex minimization, and its solutions depend upon the choice of initialization. We initialize it with the "ground-truth" partition $Y_n^i := A_i \cap X_n$ in an attempt to avoid sub-optimal solutions and to bias the algorithm towards the correct continuum cut. We terminate 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. Finally, we quantify the error between the optimal 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,

$$e_n = \min \left\{ \frac{1}{n} \sum_{i=1}^{n} |\mathbf{1}_{Y_n^i}(\mathbf{x}_i) - \mathbf{1}_{Y_n^*}(\mathbf{x}_i)|, \frac{1}{n} \sum_{i=1}^{n} |\mathbf{1}_{Y_n^i}(\mathbf{x}_i) - \mathbf{1}_{(Y_n^*)^c}(\mathbf{x}_i)| \right\}. \tag{8.3}$$

The rationale for this choice comes from the following observation. If $T_n(x)$ denotes a sequence of transportation maps between ν_n and ν that satisfy $||\operatorname{Id} - T_n||_\infty = o(1)$, then by



(a) $\varepsilon_n = n^{-0.3}$

(b) $\varepsilon_n = 2(\log n/(\pi n))^{1/2}$

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 $\mathbb{E}(\max(d_i)/\min(d_i))$, 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.

the change of variables (2.2) (and ignoring the "min" for simplicity) we have

$$e_n = \int_D |\mathbf{1}_{A_i} \circ T_n(x) - \mathbf{1}_{Y_*^n} \circ T_n(x)| \, \mathrm{d}x.$$

By the triangle inequality, we therefore obtain

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

The last inequality follows since each A_i has a piecewise smooth boundary. In this way, if $||\operatorname{Id} - T_n||_\infty = o(1)$ then verifying $e_n = o(1)$ suffices to show that TL^1 convergence of minimizers holds. Under the assumption that $||\operatorname{Id} - T_n||_\infty = o(1)$, a similar argument shows that $e_n = o(1)$ is equivalent to TL^1 convergence. This equivalence motivates using e_n as a quantitative measure of TL^1 convergence in 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

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

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. Random graph theory shows that $\mathcal{G}_n$ is almost surely connected in this regime: the probability that $\mathcal{G}_n$ is disconnected vanishes in the $n \to \infty$ limit. The second scaling also gives rise to a sequence $\mathcal{G}_n$ of connected random geometric graphs for n sufficiently large (see Gupta and Kumar (1999), Penrose (2003)). 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 $n \to \infty$, 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 bellow the connectivity threshold of random geometric graphs (see Gupta and Kumar (1999), Penrose (2003)). The graphs $\mathcal{G}_n$ are disconnected for large enough n under this scaling. However, in this regime each $\mathcal{G}_n$ has a "giant component" (a connected subgraph $\mathcal{H}_n$ of $\mathcal{G}_n$) that contains all but a small handful of vertices (see Figure 4 at right).

We designed our experiments to explore the extent to which a lack of graph-regularity or graph-connectivity might cause inconsistency 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 correctible way. As the theory outlined above indicates, if we pose the balanced 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.

Table 1 and Figure 4 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 simply due to the overwhelming computational burden. In general, we observe that sparser graphs lead to larger error (see Table 1). We caution that the corresponding rates reported in Figure 4 may not coincide with the true asymptotic rate of convergence, since we expect that as $n \to \infty$ the denser graph will still produce lower error. We furthermore remark that the measure of error we consider in Table 1 is also too weak 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 (Bresson et al., 2012), so we cannot say with certainty that the corresponding computed optimizers are global. Instead, initializing the algorithm with the "ground truth" partition biases the

$n =$	$1k$	$2k$	$4k$	$8k$	$16k$	$32k$	$64k$
$\varepsilon_n = n^{-0.3}$:							
$\mathbb{E}(e_n)$	.0776	.0616	.0495	.0391	.0320	.0238	.0205
Trials	10^4	10^4	10^4	10^4	1008	1008	192
$\varepsilon_n = 2(\log n/(\pi n))^{1/2}$:							
$\mathbb{E}(e_n)$	.0710	.0603	.0509	.0427	.0366	.0303	.0256
Trials	10^4	10^4	10^4	10^4	1008	1008	192
$\varepsilon_n = (\log n/(\pi n))^{1/2}$:							
$\mathbb{E}(e_n)$	.3221	0.1984	.1216	.0883	.0672	.0528	.0424
Trials	10^4	10^4	10^4	10^4	1008	1008	192

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 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.

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 $\mathbb{E}(e_n)$ decays to zero with a polynomial rate. The varying degree properties of the random geometric graphs in these regimes do not seem to play much of a role. A disconnected graph, while more problematic, is not an insurmountable obstacle provided $\mathcal{G}_n$ contains a giant component. A naive handling of the disconnected vertices still leads to plausibly consistent results. 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.

Acknowledgments

The authors are grateful to the editor and the referees for many valuable suggestions.

They are also 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 (grants DMS-1211760 and DMS-1516677) for its support. JvB was supported by NSF grant DMS 1312344/DMS 1521138. 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.

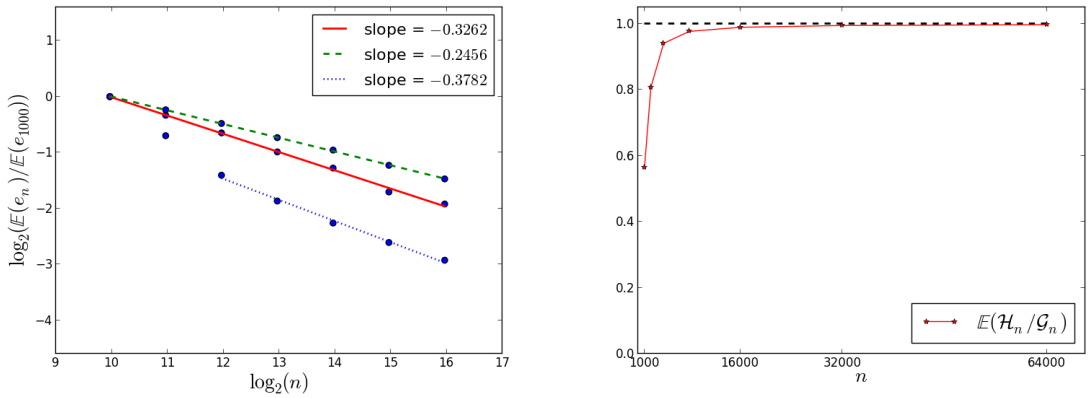


Figure 4: At left: a log-log plot of the relative expected errors $\mathbb{E}(e_n)/\mathbb{E}(e_{1000})$ computed in Table 1 together with a corresponding linear approximation for n large. The solid red line corresponds to the scaling $\varepsilon_n = n^{-0.3}$, the dashed green line corresponds to the scaling $\varepsilon_n = 2(\log n/(\pi n))^{1/2}$ and the dotted blue line corresponds to the scaling $\varepsilon_n = (\log n/(\pi n))^{1/2}$ of the disconnected regime. The linear approximation for the scaling $\varepsilon_n = (\log n/(\pi n))^{1/2}$ is given for those graphs $\mathcal{G}_n$ that, in expectation, have more than 90% of vertices in the giant component. At right: the expected fraction of vertices that lie in the giant component $\mathcal{H}_n$ of the disconnected random geometric graph $\mathcal{G}_n$.

References

G. Alberti and G. Bellettini. A non-local anisotropic model for phase transitions: asymptotic behaviour of rescaled energies. *European J. Appl. Math.*, 9(3):261–284, 1998. ISSN 0956-7925. doi: 10.1017/S0956792598003453. URL `http://dx.doi.org/10.1017/S0956792598003453`.

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. ISBN 0-19-850245-1.

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.

E. Arias-Castro and B. Pelletier. On the convergence of maximum variance unfolding. *The Journal of Machine Learning Research*, 14(1):1747–1770, 2013.

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.

S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM)*, 56(2):5, 2009.

A. Baldi. Weighted BV functions. *Houston J. Math.*, 27(3):683–705, 2001. ISSN 0362-1588.

M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. *Advances in Neural Information Processing Systems (NIPS)*, 2006.

M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *J. Comput. System Sci.*, 74(8):1289–1308, 2008. ISSN 0022-0000. doi: 10.1016/j.jcss.2007.08.006. URL `http://dx.doi.org/10.1016/j.jcss.2007.08.006`.

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. ISSN 0944-6532.

A. Braides. *Gamma-Convergence for Beginners*. Oxford Lecture Series in Mathematics and Its Applications, Oxford University Press, 2002.

A. Braides and N. K. Yip. A quantitative description of mesh dependence for the discretization of singularly perturbed nonconvex problems. *SIAM J. Numer. Anal.*, 50(4):1883–1898, 2012. ISSN 0036-1429. doi: 10.1137/110822001. URL `http://dx.doi.org/10.1137/110822001`.

X. Bresson and T. Laurent. Asymmetric Cheeger cut and application to multi-class unsupervised clustering. CAM report 12-27, UCLA, 2012.

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.

X. Bresson, T. Laurent, D. Uminsky, and J. von Brecht. Multiclass total variation clustering. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

A. Chambolle, A. Giacomini, and L. Lussardi. Continuous limits of discrete perimeters. *M2AN Math. Model. Numer. Anal.*, 44(2):207–230, 2010. ISSN 0764-583X. doi: 10.1051/m2an/2009044. URL `http://dx.doi.org/10.1051/m2an/2009044`.

J. Cheeger. A Lower Bound for the Smallest Eigenvalue of the Laplacian. *Problems in Analysis*, pages 195–199, 1970.

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.

G. Dal Maso. *An Introduction to* Γ*-convergence*. Springer, 1993.

N. Dirr and E. Orlandi. Sharp-interface limit of a Ginzburg-Landau functional with a random external field. *SIAM J. Math. Anal.*, 41(2):781–824, 2009. ISSN 0036-1410. doi: 10.1137/070684100. URL `http://dx.doi.org/10.1137/070684100`.

R. M. Dudley. *Real analysis and probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. ISBN 0-521-00754-2. doi: 10.1017/CBO9780511755347. URL http://dx.doi.org/10.1017/CBO9780511755347. Revised reprint of the 1989 original.

S. Esedoḡlu and F. Otto. Threshold dynamics for networks with arbitrary surface tensions. *Comm. Pure Appl. Math.*, 68(5):808–864, 2015. ISSN 0010-3640. doi: 10.1002/cpa.21527. URL http://dx.doi.org/10.1002/cpa.21527.

N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canad. J. Math.*, 67(6):1358–1383, 2015. ISSN 0008-414X. doi: 10.4153/CJM-2014-044-6. URL http://dx.doi.org/10.4153/CJM-2014-044-6.

N. García Trillos and D. Slepčev. Continuum limit of total variation on point clouds. *Arch. Ration. Mech. Anal.*, 220(1):193–241, 2016. ISSN 0003-9527. doi: 10.1007/s00205-015-0929-z. URL http://dx.doi.org/10.1007/s00205-015-0929-z.

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. doi: 10.1214/074921706000000888. URL http://dx.doi.org/10.1214/074921706000000888.

M. Gobbino. Finite difference approximation of the Mumford-Shah functional. *Comm. Pure Appl. Math.*, 51(2):197–228, 1998. ISSN 0010-3640. doi: 10.1002/(SICI)1097-0312(199802)51:2⟨197::AID-CPA3⟩3.3.CO;2-K. URL http://dx.doi.org/10.1002/(SICI)1097-0312(199802)51:2<197::AID-CPA3>3.3.CO;2-K.

M. Gobbino and M. G. Mora. Finite-difference approximation of free-discontinuity problems. *Proc. Roy. Soc. Edinburgh Sect. A*, 131(3):567–595, 2001. ISSN 0308-2105. doi: 10.1017/S0308210500001001. URL http://dx.doi.org/10.1017/S0308210500001001.

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. doi: 10.1145/1007352.1007441. URL http://dx.doi.org/10.1145/1007352.1007441.

P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985. ISBN 0-273-08647-2.

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.

L. Hagen and A. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Computer-Aided Design*, 11:1074 –1085, 1992.

J. Hartigan. Consistency of single linkage for high density clusters. *J. Amer. Statist. Assoc.*, 76:388–394., 1981.

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.

M. Hein and S. Setzer. Beyond Spectral Clustering - Tight Relaxations of Balanced Graph Cuts. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

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.

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

G. Leoni. *A first course in Sobolev spaces*, volume 105 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2009. ISBN 978-0-8218-4768-8.

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. ISSN 1262-3318. doi: 10.1051/ps/2012001. URL `http://www.esaim-ps.org/article_S2810012000018`.

H. Narayanan, M. Belkin, and P. Niyogi. On the relation between low density separation, spectral clustering and graph cuts. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1025–1032, 2006.

M. Penrose. A strong law for the longest edge of the minimal spanning tree. *Ann. Probab.*, 27(1):246–260, 1999. ISSN 0091-1798. doi: 10.1214/aop/1022677261. URL `http://dx.doi.org/10.1214/aop/1022677261`.

M. Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003. ISBN 0-19-850626-0. doi: 10.1093/acprof:oso/9780198506263.001.0001. URL `http://dx.doi.org/10.1093/acprof:oso/9780198506263.001.0001`.

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

A. C. Ponce. A new approach to Sobolev spaces and connections to Γ-convergence. *Calc. Var. Partial Differential Equations*, 19(3):229–255, 2004. ISSN 0944-2669. doi: 10.1007/s00526-003-0195-z. URL `http://dx.doi.org/10.1007/s00526-003-0195-z`.

O. Savin and E. Valdinoci. Γ-convergence for nonlocal phase transitions. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 29(4):479–500, 2012. ISSN 0294-1449. doi: 10.1016/j.anihpc.2012.01.006. URL `http://dx.doi.org/10.1016/j.anihpc.2012.01.006`.

J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 22(8):888–905, 2000.

A. Singer. From graph to manifold Laplacian: the convergence rate. *Appl. Comput. Harmon. Anal.*, 21(1):128–134, 2006. ISSN 1063-5203. doi: 10.1016/j.acha.2006.03.004. URL http://dx.doi.org/10.1016/j.acha.2006.03.004.

D. 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.

D. 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.

A. Szlam and X. Bresson. Total variation and Cheeger cuts. In *International Conference on Machine Learning (ICML)*, pages 1039–1046, 2010.

M. Thorpe, F. Theil, A. M. Johansen, and N. Cade. Convergence of the k-means minimization problem using Γ-convergence. *SIAM J. Appl. Math.*, 75(6):2444–2474, 2015. ISSN 0036-1399. doi: 10.1137/140974365. URL http://dx.doi.org/10.1137/140974365.

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.

Y. van Gennip and A. L. Bertozzi. Γ-convergence of graph Ginzburg-Landau functionals. *Adv. Differential Equations*, 17(11-12):1115–1180, 2012. ISSN 1079-9389.

C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, 2003. ISBN 9780821833124. URL http://books.google.com/books?id=q6kyE2ZkxrcC.

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

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. Technical Report TR 134, Max Planck Institute for Biological Cybernetics, 2004.

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

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.

H. Weyl. On the Volume of Tubes. *Amer. J. Math.*, 61(2):461–472, 1939. ISSN 0002-9327. doi: 10.2307/2371513. URL http://dx.doi.org/10.2307/2371513.

S. X. Yu and J. Shi. Multiclass spectral clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 313–319. IEEE, 2003.