SIAM J. MATH. DATA SCI.
Vol. 1, No. 1, pp. 93–120

Variational Limits of k-NN Graph-Based Functionals on Data Clouds*

Nicolas Garcia Trillos[†]

Abstract. This paper studies the large sample asymptotics of data analysis procedures based on the optimization of functionals defined on k-NN graphs on point clouds. This paper is framed in the context of minimization of balanced cut functionals, but our techniques, ideas, and results can be adapted to other functionals of relevance. We rigorously show that provided the number of neighbors in the graph $k := k_n$ scales with the number of points in the cloud as $n \gg k_n \gg \log(n)$, then with probability one the solution to the graph cut optimization problem converges towards the solution of an analogue variational problem at the continuum level.

Key words. k-NN graph, discrete to continuum limit, Gamma-convergence, spectral clustering, Cheeger cut, graph cut

AMS subject classifications. 49J55, 49J45, 60D05, 68R10, 62G20

DOI. 10.1137/18M1188999

1. Introduction.

This paper studies the large sample asymptotics of data analysis procedures based on the optimization of functionals defined on graphs; the procedures of interest include graph-based methods for clustering, classification, and semi-supervised learning. The set of vertices of the graph is a random data set $X_n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ while the edges capture the level of similarity among points. In this work our focus is on k-NN graphs, where one puts an edge between a pair of points whenever one of them is among the k-nearest neighbors of the other; we will assume from here on that X_n is a subset of Euclidean space. Our main results rigorously show that when k scales like $\log(n) \ll k \ll n$, then the solutions of the optimization problems of interest at the graph level are consistent and converge towards the solutions of analogue variational problems at the continuum level. Spectral clustering, total variation clustering, diffusion maps, and p-Laplacian regularization are all examples of graph-based data analysis procedures with a variational flavor (see [7, 9, 21, 22, 27, 28, 33, 1]), and the results and ideas in this paper can be applied to analyze their large sample limit in this k-NN setting. For expository purposes we focus on the concrete example of minimizing balance graph cuts as described below.

To get a flavor of our main results, consider for simplicity a set of random points uniformly distributed on the region D as depicted in Figure 1. A k-NN graph for a certain choice of k is then obtained as shown in Figure 2. We introduce a functional on partitions $\{A, A^c\}$ of X_n by

$$Cut_n(A, A^c) := \frac{\sum_{\mathbf{x}_i \in A} \sum_{\mathbf{x}_j \notin A} w_{ij}}{\min\{\#A, \#(X_n \setminus A)\}}, \quad A \subseteq X_n. \tag{1.1}$$

*Received by the editors May 22, 2018; accepted for publication (in revised form) October 24, 2018; published electronically February 12, 2019.

http://www.siam.org/journals/simods/1-1/M118899.html

[†]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53711 (garciatrillo@wisc.edu).

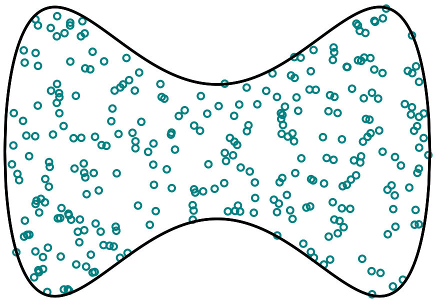

Figure 1. *A sample of $n = 120$ points.*

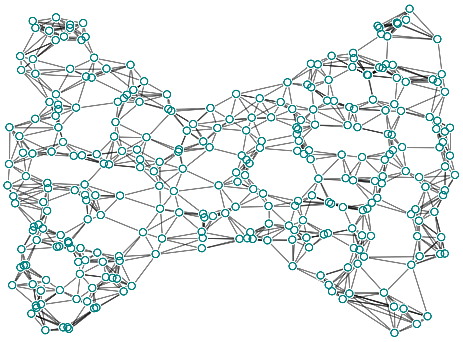

Figure 2. *Geometric graph with $k = 6$.*

The numerator favors low interaction between the two sets in a partition, while the denominator forces it to be balanced in terms of size. It is thus natural to consider the optimization problem

$$\min_{A \subseteq X_n} Cut_n(A, A^c)$$

as a sensible approach for data clustering; (1.1) is known as the *Cheeger cut* functional. In Figure 3 we illustrate the minimizer of 1.1 for the graph in Figure 2. We observe the close resemblance between the discrete minimizer in Figure 3 and the partition of the region D in Figure 4 which can be described as a solution to a variational problem at the continuum level of the form

$$\min_{A \subset D} Cut(A, A^c), \tag{1.2}$$

where the cut functional $Cut(A, A^c)$ is defined (at least for A with smooth boundary) as

$$Cut(A, A) := \frac{\int_{\partial A \cap D} dS}{\min\{|A|, |A^c|\}}.$$

The main theorem of this paper is a rigorous mathematical statement of the previous observation. More generally, we consider nonuniform densities ρ generating the data set and show that with probability one, and in a very precise sense, the solution to (1.1) converges as $n \to \infty$ towards the solution of a continuum variational problem analogue to (1.2) (but weighted appropriately in terms of the density function ρ), provided that k (the number of neighbors in the definition of the graph) scales like

$$\log(n) \ll k \ll n.$$

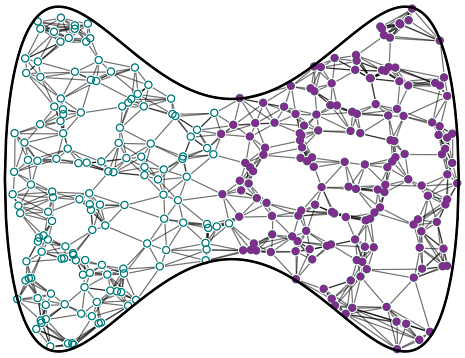

Figure 3. *Minimizer of Cheeger cut.*

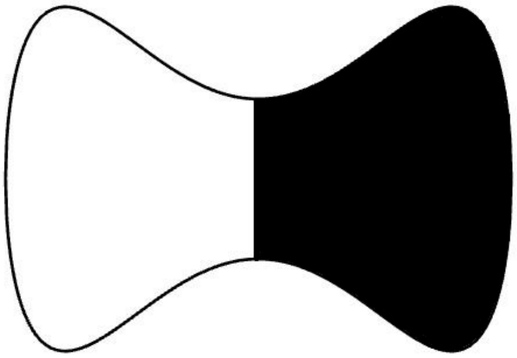

Figure 4. *Minimizer of continuous problem.*

The precise statement of our main result is presented in Theorem 2.4, the appropriate weighted continuum variational problem is defined in (2.7), and the metric under which discrete minimizers converge towards continuum minimizers, the TL^1-metric, is defined in section 2.4. Phrased in the language of data clustering, our results establish the statistical consistency of a series of clustering procedures based on minimization of graph cuts when these come from k-NN graphs. Moreover, our results show that (at least in terms of scaling with n), the condition on k needed to establish the consistency is dimension free.

To the best of our knowledge this work is the first one to rigorously address the stability of *variational* problems on k-NN graphs such as the one defined in (1.1) in the large sample limit. Most of the theoretical works found in the literature addressing similar questions assume an ε-*graph* construction on the data set (i.e., there is an edge between two points if they are within distance ε of each other); an exception to this is the work [32] where *pointwise* convergence of graph Laplacians on k-NN graphs (among other constructions) is analyzed. We find the absence of theoretical results in the k-NN setting to be a strong motivation for our work given their frequent use by practitioners due to their nicer regularity properties and their robustness to data dimensionality; see [33] for a more complete discussion on this matter. Notice that k-NN graphs can be constructed completely from ordinal information about the data points and, in particular, exact values of interpoint distances are not needed; that is, knowing the answers to all the questions of the type, is $\mathbf{x}_i$ closer to $\mathbf{x}_j$ than $\mathbf{x}_k$ is to $\mathbf{x}_l$? Or, simply in triplet form, is $\mathbf{x}_i$ closer to $\mathbf{x}_k$ than $\mathbf{x}_j$ is to $\mathbf{x}_k$? This provides enough information to construct the k-NN graph. This is a property that adds to the robustness of the k-NN construction. Figure 5 illustrates the difference between ε-graphs and k-NN graphs.

In this paper we follow the same line of thought described in [17], and, in particular, reduce our analysis to obtaining the Γ-limit (see definitions in section 2.4) of a rescaled version of the energy

$$\sum_{\mathbf{x}_i \sim_k \mathbf{x}_j} |u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)|,$$

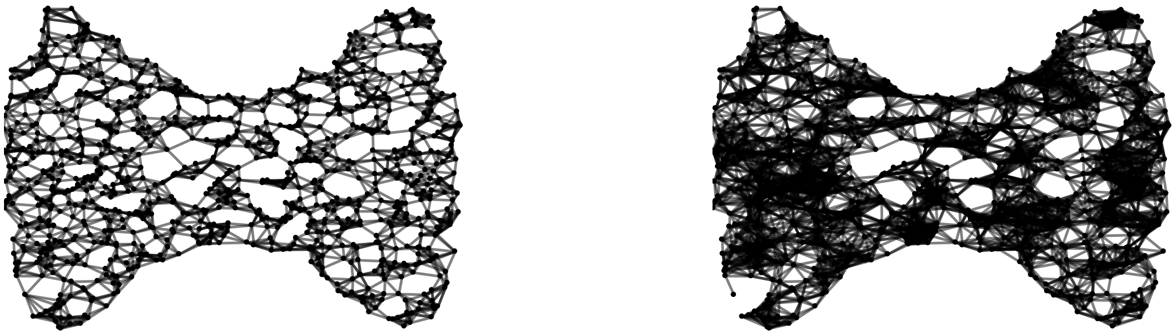

Figure 5. *On the left a k-NN graph and on the right an ε-graph. The value of ε was chosen to make sure the graph was connected. We notice the higher regularity of the k-NN graph.*

defined for functions $u_n : X_n \to \mathbb{R}$; this is the functional that appears in the numerator in (1.1) (restricting to $u = \mathbb{1}_A$). We will denote it by $GTV_{n,k}$ and refer to it as the *graph total variation*; see (2.4) for its precise definition. The Γ-limit of $GTV_{n,k}$ is shown to be a weighted (local) total variation functional at the continuum level. The notion of Γ-convergence provides precise sufficient conditions implying the stability of minimizers of variational problems; we review its definition in section 2.4 and refer the interested reader to [10] for a more complete discussion on the topic.

Most of the technical work in this paper is devoted to analyzing the "bias" of the random functional $GTV_{n,k}$. We establish the Γ-convergence of a kernel-based functional with inhomogeneous bandwidth (a sort of average of $GTV_{n,k}$) towards a (local) weighted total variation (see Propositions 3.1 and 3.2). The inhomogeneity associated with the kernel-based functional is intrinsic to the k-NN graph construction, where length-scales are determined by the Euclidean distance and the data density around each point. The fact that the resulting kernel-based approximation has a varying bandwidth makes our analysis different from that in [14]. At each point in space, the bandwidth depends inversely on the density of the ground-truth measure generating the data.

It is possible to reinterpret this feature and think of it as fixing a bandwidth but now measuring distances with an effective metric induced by the ground-truth on the ambient space; this metric, in particular, shrinks distances on regions with low density. In this work we will not pursue this idea any further, but we anticipate that there are several advantages of doing so. In particular, it is of relevance to further investigate the dimension free condition $\log(n) \ll k \ll n$, and understand better the constants appearing in these asymptotic inequalities. This is of special relevance if we want to extend our results to settings where the ground-truth distribution is, for example, a Gaussian measure in an infinite dimensional Hilbert space. Both the unboundedness of the support of the ground-truth and the infinite dimension of the ambient space are settings that are not covered by the results in this paper. We believe that a better understanding of this and other related issues are of importance for a better understanding of k-NN graphs and their benefits.

We would like to finish this introduction by mentioning some of the growing literature on large sample asymptotic analysis of operators and functionals constructed from ε-graphs. Convergence of graph Laplacians can be found in the work by Belkin and Niyogi [5], Coifman and Lafon [9], Giné and Koltchinskii [18], Hein, Audibert, and von Luxburg [20], and Singer [29]. These works deal mostly with pointwise consistency of graph Laplacians. The work of Arias-Castro, Pelletier, and Pudlo [2] studies the pointwise convergence of Cheeger energy and consequently of total variation, as well as variational convergence when the discrete functional is considered over an admissible set of characteristic functions which satisfy a "regularity" requirement. Spectral convergence of graph Laplacians, (relevant for data clustering) has been studied by, among others, Ting, Huang, and Jordan [32], Belkin and Niyogi [4], von Luxburg, Belkin, and Bousquet [34], and Singer and Wu [30]. Most of the results previously mentioned are deduced using tools from perturbation theory of linear operators. A different set of tools was introduced in [14] and later used in [12, 16, 17]. The notion of Γ-convergence (a.k.a. epi-convergence) and the introduction of a suitable metric (the TL^1-metric; see [14] and section 2.4), allowing one to compare functions at the graph level with functions at the continuum level, were crucial tools to deduce statistical consistency for a large class of balanced graph cuts [17] and for spectral clustering [16]. In all of these results, sharp convergence rates for $\varepsilon := \varepsilon_n$ guaranteeing the consistency were provided. In the context of graph-based approaches to semi-supervised learning, it is worth mentioning the work [31] where the p-Laplacian regularization (for p large enough) is studied, as well as the Bayesian approach in [13, 11] where the convergence of graph posteriors is analyzed.

1.1. Outline.

The rest of the paper is organized as follows. In section 2 we make our assumptions precise, present some of the examples of variational problems that are relevant for important tasks in data analysis, and present the main results of the paper. In section 2.4 we give the definitions of the TL^1-space and the notion of Γ-convergence; we also present some auxiliary results that are needed in what follows. Finally, in section 3 we present the proofs of the main results.

2. Set-up and main results.

We assume that the data $X_n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ are independently and identically distributed (i.i.d.) samples from a distribution ν on $\mathbb{R}^d$. ν is assumed to be an absolutely continuous measure with respect to the Lebesgue measure, with density $\rho : D \to \mathbb{R}$, where D is a bounded, connected, open set with Lipschitz boundary, and ρ is a continuous function bounded above and below by positive constants. Namely, we assume that there are positive constants $0 < \rho_{min} < \rho_{max}$ such that for every $x \in D$ we have

$$\rho_{min} \leq \rho(x) \leq \rho_{max} \quad \forall x \in D. \tag{2.1}$$

We denote by ν_n the empirical measure

$$\nu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{x}_i},$$

and write

$$\mathbf{x}_i \sim_k \mathbf{x}_j$$

whenever $\mathbf{x}_i$ is among the k-nearest neighbors of $\mathbf{x}_j$ or vice versa. More precisely, we say that $\mathbf{x}_i$ is among the k-nearest neighbors of $\mathbf{x}_j$ if $|\mathbf{x}_i - \mathbf{x}_j| \leq |\mathbf{x} - \mathbf{x}_j|$ for all $\mathbf{x}$ in a subset of X_n of cardinality $n - k + 1$.

Remark 2.1. Because ν has a density, with probability one, the condition $\mathbf{x}_i \sim_k \mathbf{x}_j$ is equivalent to

$$\nu_n\left(\overline{B(\mathbf{x}_i, r)}\right) \leq \frac{k}{n} \quad \text{or} \quad \nu_n\left(\overline{B(\mathbf{x}_j, r)}\right) \leq \frac{k}{n},$$

where $r = |\mathbf{x}_i - \mathbf{x}_j|$. Without loss of generality we assume this fact in what follows.

Let us introduce some functions that we use. Let $\eta : [0, \infty) \to \mathbb{R}$ be the step function

$$\eta(t) := \begin{cases} 1 & \text{if } t < 1, \\ 0 & \text{if } t \geq 1. \end{cases}$$

For an arbitrary $\varepsilon > 0$ we let η_ε be the function

$$\eta_\varepsilon(t) := \frac{1}{\varepsilon^d} \eta\left(\frac{t}{\varepsilon}\right), \quad t \in [0, \infty).$$

We let σ_η be the quantity

$$\sigma_\eta := \int_{\mathbb{R}^d} \eta(z)|z_1| dz, \tag{2.2}$$

where z_1 is the first coordinate of the vector z. Notice that in the above expression z_1 can be replaced with $z \cdot v$ for any vector v with unit norm.

2.1. Graph total variation and total variation in the continuum.

In order to introduce the *graph total variation* functional $GTV_{n,k}$ (an appropriate rescaled version of the numerator in (1.1)), it will be convenient to define

$$J_k(\mathbf{x}_i, \mathbf{x}_j) := \begin{cases} 1, & \mathbf{x}_i \sim_k \mathbf{x}_j, \\ 0 & \text{otherwise}, \end{cases}$$

and for $k := k_n \in \mathbb{N}$, let $\bar{\varepsilon}_n$ be the number for which

$$\bar{\varepsilon}_n^d = \frac{k_n}{n}. \tag{2.3}$$

The *graph total variation* functional is then defined as

$$GTV_{n,k}(u_n) := \frac{1}{n^2 \bar{\varepsilon}_n^{d+1}} \sum_{i,j} J_k(\mathbf{x}_i, \mathbf{x}_j)|u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)|, \quad u_n \in L^1(\nu_n). \tag{2.4}$$

Notice that when $u_n = \mathbb{1}_{A_n}$ for some $A_n \subseteq X_n$, $GTV_{n,k}(u_n)$ is just a rescaled version of the numerator of the Cut_n functional in (1.1). At the discrete level we are interested in the following optimization problem:

$$\min_{A_n \subseteq X_n} \frac{GTV_{n,k}(\mathbb{1}_{A_n})}{\min\{\nu_n(A_n), \nu_n(X_n \setminus A_n)\}}. \tag{2.5}$$

The continuum counterpart of the graph total variation is a functional that takes the following form.

Definition 2.2. *Let $h : D \to \mathbb{R}$ be a continuous function bounded below and above by positive constants. For an arbitrary function $u \in L^1(D)$, we define its weighted (by h) total variation as*

$$TV(u;h) := \sup \left\{ \int_D u(x)\,\mathrm{div}(\zeta) dx \ : \ \zeta \in C_c^\infty(D : \mathbb{R}^d), \quad |\zeta(x)| \leq h(x) \quad \forall x \in D \right\}. \tag{2.6}$$

In the above and in the remainder of the paper, we use $L^1(D)$ to represent the space of L^1 functions with respect to the Lebesgue measure restricted to D. In addition, when $h \equiv 1$ we write $TV(u)$ instead of $TV(u;h)$ and we denote by $BV(D)$ the space of functions $u \in L^1(D)$ for which $TV(u) < \infty$. Notice that for any h continuous and bounded above and below by positive constants, the condition $TV(u) < \infty$ is equivalent to the condition $TV(u;h) < \infty$.

Remark 2.3. If $u \in BV(D)$ is smooth, then

$$TV(u;h) = \int_D |\nabla u| h(x) dx.$$

Also, if $u = \mathbb{1}_A$ for some open set A with smooth boundary, then

$$TV(u;h) = \int_{\partial A \cap D} h(x) d\mathcal{H}^{d-1}(x),$$

where $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure.

We will be interested in a variational problem of the form

$$\min_{A \subseteq D} \frac{TV(\mathbb{1}_A;h)}{\min\{\nu(A), \nu(D \setminus A)\}} \tag{2.7}$$

for an appropriately chosen function h. We can motivate the form of the numerator in the variational problem (2.7) as follows. For a given $u : D \to \mathbb{R}$ sufficiently smooth we have

$$\mathbb{E}(GTV_{n,k}(u)) \sim \left(\frac{n}{k}\right)^{1+1/d} \int_D \int_{B(x,\varepsilon_n(x))} |u(x) - u(y)| \rho(x)\rho(y) dy dx,$$

where $\varepsilon(x)$ is the radius needed for the ball $B(x, \varepsilon(x))$ to contain the k-nearest neighbors of x in the point cloud. This radius is such that

$$\rho(x)(\varepsilon_n(x))^d \sim \frac{k}{n}.$$

On the other hand, when k is small compared to n, $\varepsilon_n(x)$ is small and thus, for $y \in B(x, \varepsilon_n(x))$, the term $|u(y)-u(x)|$ is approximated by $|\nabla u(x) \cdot (y-x)|$ which in turn is of order $\varepsilon(x)|\nabla u(x)|$. Moreover, for such a y the value of $\rho(y)$ is close to the value of $\rho(x)$ (using the smoothness of the density) so that $\rho(y)\rho(x) \sim \rho(x)^2$. From these estimates we see that

$$\mathbb{E}(GTV_{n,k}(u)) \sim \int_D |\nabla u(x)| \rho^{1-1/d} dx.$$

That is, $h = \rho^{1-1/d}$.

The previous computations only serve as motivation for our results, and we must emphasize that they do not constitute a proof (nor a sketch of proof) of our main results. We recall that our interest is in studying the *variational* convergence of discrete energies, and not their *pointwise* convergence.

2.2. Main results. Our main result establishes the convergence of minimizers of (2.5) towards minimizers of (2.7). Notice, however, that solutions to (2.5) are discrete sets, whereas solutions to (2.7) are continuum sets, and so we need to clarify the sense in which we will establish the convergence of discrete minimizers towards continuum ones. The convergence is taken in the TL^1-space introduced in [14], where, in particular, functions on the point cloud and functions on D are seen as elements of the same space (see section 2.4 for details).

We are ready to establish our main result.

Theorem 2.4. *Let $d \geq 3$ and let $D \subseteq \mathbb{R}^d$ be an open, connected, and bounded domain with Lipschitz boundary. Let $\rho : D \to \mathbb{R}$ be a continuous density function satisfying (2.1) and let ν be the probability measure $d\nu = \rho dx$. Let $\{k_n\}_{n\in\mathbb{N}}$ be a sequence of natural numbers satisfying*

$$\lim_{n\to\infty} \frac{k_n}{\log(n)} = +\infty, \quad \lim_{n\to\infty} \frac{k_n}{n} = 0.$$

Let $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be i.i.d. points from ν and let A_n^* be a solution to (2.5).

Then, with probability one, along every subsequence of $\{A_n^*\}_{n\in\mathbb{N}}$ there is a further subsequence converging in the TL^1-sense towards a minimizer of (2.7), where

$$h(x) := \rho^{1-1/d}(x).$$

Moreover, the minimum value of (2.5) converges as n goes to infinity towards $\sigma_\eta/\alpha_d^{1+1/d}$ times the minimum value of (2.7), where σ_η is defined in (2.2) and α_d is the volume of the unit ball in $\mathbb{R}^d$.

Remark 2.5. In the above theorem if the minimizer $\{A^*, A^{*c}\}$ of (2.7) is unique, then the convergence is along the entire sequence of discrete minimizers.

As discussed in the introduction, and especially as shown in [17], our results are a direct corollary of Theorems 2.6 and 2.7 below. We show that the Γ-limit (in the TL^1-sense) of the functional GTV_{n,k_n} (for the same scaling $k = k_n$ as in Theorem 2.4) is the functional $TV(\cdot; \rho^{1-1/d})$. The notion of Γ-convergence and its connection to the stability of minimizers of functionals are reviewed in section 2.4.

Theorem 2.6 (Γ-convergence). *Let $d \geq 3$ and let $D \subseteq \mathbb{R}^d$ be an open, connected, and bounded domain with Lipschitz boundary. Let $\rho : D \to \mathbb{R}$ be a continuous density function satisfying (2.1). Let $\{k_n\}_{n\in\mathbb{N}}$ be a sequence of natural numbers satisfying*

$$\lim_{n\to\infty} \frac{k_n}{\log(n)} = +\infty, \quad \lim_{n\to\infty} \frac{k_n}{n} = 0.$$

Then, the functionals GTV_{n,k_n} Γ-converge towards $\frac{\sigma_\eta}{\alpha_d^{1+1/d}} TV(\cdot; \rho^{1-1/d})$ in the TL^1-sense, where σ_η is defined in (2.2), and α_d is the volume of the unit ball in $\mathbb{R}^d$. Furthermore, when restricted to indicator functions, the energies GTV_{n,k_n} Γ-converge to the functional $\frac{\sigma_\eta}{\alpha_d^{1+1/d}} TV(\cdot; \rho^{1-1/d})$ restricted to indicator functions.

Theorem 2.7 (compactness). *Under the same assumptions in Theorem 2.6, the following statement holds with probability one: If $\{u_n\}_{n\in\mathbb{N}}$ is a sequence with $u_n \in L^1(\nu_n)$ for which*

$$\sup_{n\to\infty} \|u_n\|_{L^1(\nu_n)} < \infty$$

and

$$\sup_{n\to\infty} GTV_{n,k_n}(u_n) < \infty,$$

then $\{u_n\}_{n\in\mathbb{N}}$ is precompact in $TL^1(D)$. That is, every subsequence of $\{u_n\}_{n\in\mathbb{N}}$ has a further subsequence converging in $TL^1(D)$.

Remark 2.8. The above theorems hold for domains D in $\mathbb{R}^2$ provided we replace the condition $\lim_{n\to\infty} \frac{k_n}{\log(n)} = \infty$ with the condition $\lim_{n\to\infty} \frac{k_n}{(\log(n))^{3/2}} = \infty$. This can be seen directly from our proofs, and follows from the fact that the rate of convergence of the ∞-transportation distance between ν and ν_n (for $d = 2$) scales like $\frac{\log(n)^{3/4}}{n^{1/2}}$ and not like $\frac{(\log(n))^{1/2}}{n^{1/2}}$ (see [15]).

Remark 2.9. We would like to remark that Theorems 2.4, 2.6, and 2.7 continue to be true if the graph construction is slightly changed. For example, suppose that the graph total variation is defined using a nonsymmetric version of the k-NN graph:

$$J_{k,a}(\mathbf{x}_i, \mathbf{x}_j) := \begin{cases} 1 & \text{if } \mathbf{x}_i \sim_{k,a} \mathbf{x}_j, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\mathbf{x}_i \sim_{k,a} \mathbf{x}_j,$$

whenever $\mathbf{x}_j$ is among the k-nearest neighbors of $\mathbf{x}_i$. Then, the continuum limit of the resulting graph total variation is still $\frac{\sigma_\eta}{\alpha_d^{1+1/d}} TV(\cdot;\rho^{1-1/d})$ as in our main theorems. No modification to the proofs is actually needed to see this. We also notice that the symmetrization with weights,

$$J_{k,s}(\mathbf{x}_i, \mathbf{x}_j) := \frac{1}{2} J_{k,a}(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2} J_{k,a}(\mathbf{x}_j, \mathbf{x}_i),$$

produces the same graph total variation as the nonsymmetric k-NN graph. The bottom line is that, at least asymptotically, all three constructions of k-NN graphs (the previous two and the one explicitly considered in our results) give rise to the same associated variational problem in the limit.

Remark 2.10. The scaling

$$\log(n) \ll k_n \ll n$$

for k_n appearing in our main results has been used by practitioners and can be motivated by heuristic arguments. Indeed, at least in terms of scaling, a k-NN graph can be related to an ε-graph (two points are connected with an edge if they are within distance ε of each other) by

$$\frac{k}{n} \sim \varepsilon^m.$$

The results on the connectivity of random geometric graphs (i.e., ε-graphs; see [25]) imply that in order to get a connected graph (asymptotically), ε must scale with n like

$$\frac{\log(n)^{1/m}}{n^{1/m}} \lesssim \varepsilon_n$$

which translates into a condition for k_n:

$$\log(n) \lesssim k_n.$$

In particular, we do not expect to see convergence of discrete minimizers of (2.5) towards minimizers of functionals like (2.7) when k_n scales like $k_n \ll \log(n)$. Indeed, in such a scenario discrete minimizers have (asymptotically) zero energy, while continuum minimizers do not. From that point of view, at least when $m \geq 3$, our results are sharp.

Remark 2.11. The above set-up and results can be extended to the setting in which the support of ν is not a domain D contained in the ambient space $\mathbb{R}^d$ but a compact manifold $\mathcal{M} \subseteq \mathbb{R}^d$ with intrinsic dimension m. When that is the case, the appropriate scalings for the functionals are obtained using the intrinsic dimension m and not the dimension of the ambient space d. Although in this paper we omit the details of such extension, we point out that the results in [26] (later used in the proofs in [14]) can be adapted to the manifold setting by establishing results analogous to those in [15] in the manifold setting, using the geodesic flow in $\mathcal{M}$ and the fact that the Euclidean distance (distance in the ambient space $\mathbb{R}^d$) is a third order approximation for the geodesic distance (intrinsic distance), i.e.,

$$d_{\mathcal{M}}(x,y) = |x-y| + O(|x-y|^3), \quad x, y \in \mathcal{M}.$$

The details can be presented elsewhere in a more general setting which is also of interest; see Remark 2.12 below.

Remark 2.12. Suppose that $d \geq 3$. One of the important consequences of Theorem 2.6 is that the admissible regimes for $k = k_n$ that guarantee the recovery of a nontrivial variational limit for the energies $GTV_{n,k}$ do not depend on d (we impose $n \gg k_n \gg \log(n)$). As a consequence, notice that despite the fact that the scaling factor appearing in $GTV_{n,k}$ depends on d, the minimizers of the Cheeger cut are unaffected by a rescaling of the energy by a positive constant. In other words, in principle we do not need to know the dimensionality of the data beforehand in order to obtain good clusters using the graph total variation associated to a k-NN graph.

2.3. Other examples of relevant variational problems on graphs.

In this section we present a variety of examples of other optimization problems on graphs that are relevant for data analysis. The common structure in all of these optimization problems is that in their objective functions the highest order term is either a graph total variation or an L^p version of it; for this reason Theorems 2.6 and 2.7 (in conjunction with Proposition 2.21) are of relevance beyond the setting of Theorem 2.4.

For the rest of this section $(w_{ij})_{ij}$ represents a similarity graph on a data set (not necessarily a k-NN graph).

Example 2.13 (ratio graph cuts for clustering). A functional closely related to the Cheeger cut in (1.1) is the *ratio cut* defined by

$$Cut_R(A, A^c) := \frac{\sum_{\mathbf{x}_i \in A} \sum_{\mathbf{x}_j \notin A} w_{ij}}{|A| \cdot |A^c|}, \quad A \subseteq X_n.$$

Both the above functional and the Cheeger cut can be seen as examples of a larger class of functionals known as *balanced cuts*. These types of functionals penalize partitions $\{A, A^c\}$ of X_n that either have a big interface separating A and A^c or are highly unbalanced in the sense that A and A^c are of dissimilar size; intuitively, these are desirable features for a good partitioning of the data and motivate considering the minimization problem

$$\min_{A \subseteq X_n} Cut(A, A^c)$$

to obtain a good partition of X_n. Notice that the numerator in both functionals is, up to a multiplicative factor, the graph total variation defined by

$$GTV(\mathbb{1}_A) := \sum_{i,j} w_{ij} |\mathbb{1}_A(\mathbf{x}_i) - \mathbb{1}_A(\mathbf{x}_j)|.$$

Assuming that the set of vertices consists of samples from the distribution ν and that the weights of the graph are those obtained from an ε-graph (with $\varepsilon := \varepsilon_n$ chosen appropriately), the results in [17] state that minimizers of the Cheeger and ratio cuts converge, as $n \to \infty$, towards solutions to analogue variational problems in the continuum. This result can then be interpreted as a consistency result for clustering using balanced cuts.

Example 2.14 (graph total variation in the context of classification). Suppose that

$$\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$$

are samples from some distribution $\boldsymbol{\nu}$ supported on $\overline{D} \times \{0, 1\}$. The pair $(\mathbf{x}_i, y_i)$ is interpreted as the feature values ($\mathbf{x}$ variable) and label (y variable) of an individual from a given population. Based on the observed data (training data) the idea is to construct a "good" *classifier* assigning labels to every potential individual in the population. A typical choice of risk functional used to define "good" classifiers is the *average misclassification error*. Unfortunately, when the distribution $\boldsymbol{\nu}$ is unknown (as is usually the case) it is not possible to determine the classifier that minimizes the average misclassification error (Bayes classifier), and hence an approximation to it based exclusively on the observed data is the best thing one can hope for. Such an approximation can be obtained using the graph total variation as we describe below.

Given the weighted graph $(\{\mathbf{x}_i\}_i, (w_{ij})_{i,j})$, consider the energy

$$R_{n,\lambda}(u) := R_n(u) + \lambda GTV(u), \quad u : \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \to \mathbb{R}.$$

The functional R_n is the *empirical risk*, and the parameter λ is introduced so as to emphasize or deemphasize the regularizing effect of the graph total variation. In [12], the weights for the graph are assumed to be those coming from an ε-graph, and the problem of obtaining an approximation to the Bayes classifier is divided into first solving the minimization problem

$$\min_{u : X_n \to \mathbb{R}} R_{n,\lambda}(u),$$

and then extending the minimizer to the whole ambient space appropriately. With the right scaling for $\lambda = \lambda_n$, one can establish the asymptotic consistency of the constructed approximation. See [12] for more details.

Example 2.15 (spectral clustering and spectral embeddings). Undoubtedly, one of the most popular graph-based methods for data clustering is *spectral clustering*. In the two way clustering setting, spectral clustering can be seen as a relaxation of the ratio cut minimization problem mentioned in Example 2.13. Indeed, the relaxed problem takes the form

$$\min_{u:X_n\to\mathbb{R}} \frac{\sum_{i,j} w_{ij}(u(\mathbf{x}_i)-u(\mathbf{x}_j))^2}{\sum_i (u(\mathbf{x}_i)-\overline{u})^2},$$

where $\overline{u}$ is the average $\overline{u} := \frac{1}{n}\sum_i u(v_i)$; the numerator of the above objective function is an L^2-version of the graph total variation. It is well known that the above optimization problem is actually an eigenvalue problem for the graph Laplacian and that the first nontrivial eigenvector of the graph Laplacian is a minimizer. In the context of multiway clustering, higher eigenvectors of the graph Laplacian are used to define an embedding of the data cloud into a Euclidean space with low dimension. In turn, the embedded data can be clustered using an algorithm like k-means, inducing in this way a partition for the original data. Among the many results in the literature addressing the consistency of spectral clustering (see the introduction for some references), we highlight the work in [16] which exploits the variational characterization of eigenvectors/eigenvalues. In that work the graph on the cloud X_n is assumed to be an ε-graph.

Example 2.16 (p-Laplacian regularization for semi-supervised learning). Let

$$\{(\mathbf{x}_1,y_1),\dots,(\mathbf{x}_q,y_q)\}$$

be q labeled data points and let

$$\mathbf{x}_{q+1},\dots,\mathbf{x}_n$$

be a set of unlabeled data points. We think of n as being much larger than q.

In [1] the authors study the optimization problem

$$\min_{u:X_n\to\mathbb{R}} \sum_{i,j} w_{ij}|u(\mathbf{x}_i)-u(\mathbf{x}_j)|^p, \tag{2.8}$$

subject to $u(\mathbf{x}_i) = y_i$, $i = 1,\dots,q$. The above problem is used to construct a regressor in the context of semi-supervised learning; here p is a user chosen parameter, whose role is to impose regularity on candidate functions u (the higher p is, the more regular the functions will be). In two independent contributions [8, 31] the authors study the large data limit of this variational problem and by studying the discrete regularity induced by the L^p-term in (2.8), they are able to rigorously show that there is a phase transition in the value of p at which solutions to (2.8) "stop forgetting" the q labeled data points; the transition occurs at $p > m$ (where m is the intrinsic dimension of the x data set), a result that is reminiscent of the Sobolev embedding theorem. The setting in which these results are shown is that of ε-graphs.

Example 2.17 (Bayesian formulation of semi-supervised learning). In the context of Example 2.16, a related optimization problem is

$$\min_{u:X_n\to\mathbb{R}} \langle \Delta_n^\alpha u, u\rangle + \frac{1}{2\gamma^2}\sum_{i=1}^{q} |u(\mathbf{x}_i) - y_i|^2,$$

where Δ_n is the graph Laplacian associated to the graph $(w_{ij})_{ij}$ and where α is a positive number whose role is to enforce more or less regularity on a candidate function (higher α results in more regularity). The minimizer of this functional can be seen as the MAP of the *posterior* distribution

$$p(u|y) \propto \exp(-\phi(u;y))d\pi(u)$$

where π is a *prior* distribution (in this case Gaussian with covariance matrix $\Delta_n^{-\alpha}$) and ϕ is a *negative log-likelihood* model (in this case additive Gaussian noise).

This Bayesian point of view to graph-based semi-supervised learning was introduced in [6]. In [13] the authors study the passage to the large $n \to \infty$ limit (q fixed) and study the consistency of posterior measures. Moreover, from said consistency result and from the properties of the limiting posterior distribution, the authors in [11] provide some theory supporting the MCMC algorithm to sample from the posterior $p(u|y)$ that was proposed in [6] . This algorithm was introduced so as to alleviate the curse of dimensionality when sampling from $p(u|y)$ (here the curse of dimensionality arises from the large number n). The setting in which all of these results are shown is that of ε-graphs.

In view of the results presented in this paper we can study, in a straightforward way, the consistency of solutions to optimization problems like the ones in the previous examples in the k-NN graph setting.

2.4. Preliminaries.

The purpose of this subsection is to present some definitions and preliminary results that we use in the proof of our main theorems. In particular, we present the definitions of TL^1-space and Γ-convergence.

Definition 2.18. *The space $TL^1(D)$ is defined as the set of all pairs (μ, u) where μ is a Borel probability measure on D and u is a function in $L^1(D,\mu)$ (written $L^1(\mu)$ from now on). This set can be endowed with the TL^1-metric defined by*

$$d_{TL^1}((\mu_1,u_1),(\mu_2,u_2)) := \inf_{\pi\in\Gamma(\mu_1,\mu_2)} \left\{ \int_{D\times D} |x-y| d\pi(x,y) + \int_{D\times D} |u_1(x) - u_2(y)| d\pi(x,y) \right\},$$

where $\Gamma(\mu_1,\mu_2)$ stands for the set of couplings between μ_1 and μ_2, that is, the set of all probability measures on the product $D\times D$ whose first and second marginals are μ_1 and μ_2, respectively.

Given that with probability one the empirical measure ν_n converges weakly to ν as $n\to\infty$, and given that ν has a density, we may use the characterization of TL^1-convergence from Proposition 3.12 in [14] to conclude that $\{(\nu_n,u_n)\}_{n\in\mathbb{N}}$ converges to (ν,u) in TL^1 if and only if *there exists* a sequence of transportation maps $\{T_n\}_{n\in\mathbb{N}}$ with $T_{n\sharp}\nu = \nu_n$ (T_n pushes forward ν into ν_n), satisfying

$$\lim_{n\to\infty} \int_D |T_n(x) - x| d\nu(x) = 0 \tag{2.9}$$

and

$$\lim_{n\to\infty}\int_D |u_n(T_n(x)) - u(x)| d\nu(x) = 0. \tag{2.10}$$

In turn, this holds if and only if *for all* sequences of transportation maps with $T_{n\sharp}\nu = \nu_n$ that satisfy (2.9) one has (2.10). Because of this characterization, we may abuse notation slightly and simply write $u_n \xrightarrow{TL^1} u$ without specifying the corresponding attached measures whenever it is clear from the context that they can be omitted. Moreover, if A_n is a subset of X_n and A is a measurable subset of D, we say that A_n converges in TL^1 to A when the corresponding indicator functions converge in TL^1; this is the type of convergence that we use to establish the stability of minimizers in Theorem 2.4.

A special choice of transportation maps between ν and ν_n that we use in the remainder is provided by [15]. We may consider transportation maps $T_n : D \to X_n$ between the measures ν and ν_n satisfying the condition

$$\|Id - T_n\|_\infty \leq \frac{C(\log(n))^{1/d}}{n^{1/d}} =: \delta_n \tag{2.11}$$

for some constant C (provided $d \geq 3$). Indeed, the results in [15] show that with probability one, there exist transportation maps $\{T_n\}_{n\in\mathbb{N}}$ for which $T_{n\sharp}\nu = \nu_n$ and for which (2.11) holds for all large enough $n \in \mathbb{N}$. Since all of our results are asymptotic in nature, we may as well assume for the remainder that with probability one (2.11) holds for all $n \in \mathbb{N}$. One last relevant property of the map T_n, which follows directly from the fact that it transports ν into ν_n, is the change of variables formula

$$\int_D f(x) d\nu_n(x) = \int_D f(T_n(x)) d\nu(x), \tag{2.12}$$

which allows us to write integrals with respect to ν_n in terms of integrals with respect to ν.

We now present the definition of Γ-convergence in the context of a general metric space. This is a notion of convergence for functionals which together with a coercivity assumption guarantees the stability of minimizers in the limit. A standard reference for Γ-convergence is [10].

Definition 2.19. *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and let $F_n : \mathcal{X} \to [0,\infty]$ be a sequence of functionals. The sequence $\{F_n\}_{n\in\mathbb{N}}$ Γ-converges with respect to the metric $d_{\mathcal{X}}$ to the functional $F : \mathcal{X} \to [0,\infty]$ as $n \to \infty$ if the following properties hold:*

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

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

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

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

- Compactness: *Every bounded sequence* $\{x_n\}_{n\in\mathbb{N}}$ *satisfying*
$$\sup_{n\in\mathbb{N}} F_n(x_n) < \infty$$
is precompact.

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

Remark 2.20. A sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathcal{X}$ like the one appearing in the limsup inequality is said to be a recovery sequence for x. It is straightforward to show that if one can find recovery sequences for all elements in a set $\mathcal{X}'$ satisfying

1. For all $x \in \mathcal{X}$ there exists a sequence $\{x_l\}_{l\in\mathbb{N}} \subseteq \mathcal{X}'$ such that $x_l \to x$ and $F(x_l) \to F(x)$,

then one can find recovery sequences for all elements in $\mathcal{X}$; such a set $\mathcal{X}'$ is said to be dense in $\mathcal{X}$ with respect to F. This fact follows from a simple diagonal argument (see [10]).

The most relevant property of Γ-convergence (in particular, for our purposes) is presented in the following proposition which can be found in [10].

Proposition 2.21. *Let $(\mathcal{X}, d_{\mathcal{X}})$ be a metric space and let $F_n : \mathcal{X} \to [0,\infty]$ be a sequence of functionals that are not identically equal to ∞. If $\{F_n\}_{n\in\mathbb{N}}$ Γ-converges towards F, and assuming the compactness property in Definition* 2.19 *holds, then we have the following:*

1. *Any sequence $\{x_n^*\}_{n\in\mathbb{N}}$ where x_n^* is a minimizer of F_n is precompact, and each of its accumulation points is a minimizer of F. In particular, if F has a unique minimizer, then $\{x_n^*\}_{n\in\mathbb{N}}$ converges towards it.*
2. *We have*
$$\lim_{n\to\infty} \min_{x\in\mathcal{X}} F_n(x) = \min_{x\in\mathcal{X}} F(x).$$

We have presented the notion of Γ-convergence in the above generality because some of the Γ-limits we will consider in this paper are taken in the context of the metric space $L^1(D)$, whereas others are taken in the context of the metric space $TL^1(D)$. Also, notice that when the functionals are allowed to be random (as is the case for the graph total variation), Γ-convergence has to be interpreted as in Definition 2.11 in [14], that is, with probability one the statement of the liminf inequality, limsup inequality, and compactness holds.

To conclude this section we list two important properties of the weighted total variation functional $TV(\cdot;h)$. Let $h : D \to \mathbb{R}$ be a continuous function which is bounded above and below by positive constants. The first property is a representation formula for $TV(u;h)$ in terms of the distributional derivative of u. Indeed, it follows from the work in [3] that for every $u \in BV(D)$ one can write

$$TV(u;h) = \int_D h(x)d|Du|(x), \tag{2.13}$$

where in the above Du stands for the distributional derivative of u (which, in general, is a signed measure) and $|Du|$ stands for the total variation measure associated to Du.

The second property that is relevant for our purposes is the *coarea formula* which states that for every $u \in BV(D)$,

$$TV(u;h) = \int_{-\infty}^{\infty} TV(\mathbb{1}_{\{x\,:\,u(x)>t\}};h)dt. \tag{2.14}$$

This formula says that the total variation of u can be written in terms of the total variation of its level sets. See Theorem 13.25 in [23].

3. Proofs of theorems.

The motivation given at the end of section 2.1 suggests that an energy of the form

$$\left(\frac{n}{k}\right)^{1+1/d}\int_D\int_{B(x,\varepsilon_n(x))}|u(x)-u(y)|\rho(x)\rho(y)dydx, \quad u\in L^1(D),$$

plays an important role in our analysis. This is indeed the case as shown in the following proposition, where we relate this nonlocal continuum energy with the local energy $TV(\cdot;\rho^{1-1/d})$. In anticipation of the next steps in the proof of our main results we introduce two length-scales $\varepsilon_n(x)$ and $\hat{\varepsilon}_n(x)$ that we assume are uniformly and asymptotically equivalent. There are no random objects in the following result, and we believe it is fair to say that this corresponds to studying the "bias" in our problem.

Proposition 3.1. *For every $n\in\mathbb{N}$ let $\varepsilon_n,\hat{\varepsilon}_n : D\to(0,\infty)$ be functions satisfying*

$$\sup_{x\in D}\left|\frac{\varepsilon_n(x)}{\hat{\varepsilon}_n(x)}-1\right|\to 0 \quad \text{as } n\to\infty$$

and

$$\sup_{x\in D}\varepsilon_n(x)\to 0 \quad \text{as } n\to\infty.$$

For $n\in\mathbb{N}$ consider the energy $F_n : L^1(D)\to[0,\infty)$ defined by

$$F_n(u) := \int_D \hat{f}_n(x)\left(\int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|)|u(x)-u(y)|\rho(y)dy\right)\rho(x)dx, \quad u\in L^1(\nu),$$

where

$$\hat{f}_n(x) := \frac{(\hat{\varepsilon}_n(x))^d}{(\nu(B(x,\varepsilon_n(x))))^{1+1/d}}.$$

Then, as $n\to\infty$, F_n Γ-converges in the $L^1(D)$-sense towards F, where F is given by

$$F(u) := \frac{\sigma_\eta}{\alpha_d^{1+1/d}}TV(u;\rho^{1-1/d}).$$

Proposition 3.1 is a consequence of the next lemma.

Lemma 3.2. *Let $\rho_1,\rho_2 : \mathbb{R}^d\to(0,\infty)$ be two Lipschitz continuous functions which are bounded above and below by positive constants in D. Let μ be the Borel measure on $\mathbb{R}^d$ given by $d\mu(x)=\rho_2(x)dx$. Let $\varepsilon_n,\hat{\varepsilon}_n : D\to(0,\infty)$ be functions satisfying*

$$\sup_{x\in D}\left|\frac{\varepsilon_n(x)}{\hat{\varepsilon}_n(x)}-1\right|\to 0 \quad \text{as } n\to\infty$$

and

$$\sup_{x\in D}\varepsilon_n(x)\to 0 \quad \text{as } n\to\infty.$$

Let $F_n : L^1(D) \to [0,\infty)$ *be defined by*

$$F_n(u) := \int_D \hat{f}_n(x) \left(\int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|)|u(x)-u(y)|\rho_1(y)dy \right) \rho_1(x)dx, \quad u \in L^1(\nu),$$

where

$$\hat{f}_n(x) := \frac{(\hat{\varepsilon}_n(x))^d}{(\mu(B(x,\varepsilon_n(x))))^{1+1/d}}.$$

Then, as $n \to \infty$, F_n Γ*-converges in the* $L^1(D)$ *sense towards*

$$F(u) := \frac{\sigma_\eta}{\alpha_d^{1+1/d}} TV\left(u; \frac{\rho_1^2}{\rho_2^{1+1/d}}\right).$$

Proof. Let us start by proving the liminf inequality. That is, let us prove that for every $u \in L^1(D)$ and every sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq L^1(D)$ satisfying $u_n \overset{L^1(D)}{\longrightarrow} u$, we have

$$\liminf_{n\to\infty} F_n(u_n) \geq F(u). \tag{3.1}$$

The following simplifications are standard. First, working along subsequences we may assume without loss of generality that the liminf is actually a limit. In addition, we may assume that both of the terms involved in inequality (3.1) are finite. We now split the proof of (3.1) into several steps.

Step 1. Instead of working with the energy F_n directly, we first consider a simpler related energy E_n defined by

$$E_n(v) := \int_D \left(\frac{1}{\hat{\varepsilon}_n(x)} \int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|)|v(x)-v(y)|dy \right) g(x)dx, \quad v \in L^1(D),$$

where

$$g(x) := \frac{(\rho_1(x))^2}{(\alpha_d \rho_2(x))^{1+1/d}}, \quad x \in D.$$

Notice that for every $x \in D$ we have

$$\left| \mu(B(x,\hat{\varepsilon}_n(x))) - \alpha_d \rho_2(x)(\hat{\varepsilon}_n(x))^d \right| \leq \alpha_d \operatorname{Lip}(\rho_2)(\hat{\varepsilon}_n(x))^{d+1}.$$

From the previous inequality, the Lipschitz continuity of ρ_1, ρ_2, and the assumptions on $\hat{\varepsilon}_n(\cdot)$ and $\varepsilon_n(\cdot)$, we deduce that for every $x \in D$ and $y \in B(x,\hat{\varepsilon}_n(x))$,

$$\left| \frac{g(x)}{\hat{\varepsilon}_n(x)} - \hat{f}_n(x)\rho_1(x)\rho_1(y) \right| \leq C \sup_{z\in D} \varepsilon_n(z),$$

where C is a constant that does not depend on x or y. It then follows that

$$\lim_{n\to\infty} |F_n(u_n) - E_n(u_n)| = 0.$$

In particular, to obtain (3.1) we may as well show that

$$\liminf_{n\to\infty} E_n(u_n) \geq F(u).$$

Step 2. Let B be a closed ball contained in D. We claim that

$$\liminf_{n\to\infty} \int_B \left(\frac{1}{\hat{\varepsilon}_n(x)} \int_B \eta_{\hat{\varepsilon}_n(x)}(|x-y|)|u_n(x)-u_n(y)|dy \right) dx \geq \sigma_\eta TV_B(u),$$

where $TV_B(u)$ is defined as

$$TV_B(u) := \sup\left\{ \int_{B^\circ} u(x)\operatorname{div}(\zeta)dx \ : \ \zeta \in C_c^\infty(B^\circ : \mathbb{R}^d), \quad |\zeta(x)| \leq 1 \quad \forall x \in B^\circ \right\}, \tag{3.2}$$

where in the above B° stands for the interior of the closed ball B (the ball without its boundary).

The claim follows from Theorem 8 in [26]. The only difference between the setting we consider here and the setting considered in [26] is that in our case the length-scale $\hat{\varepsilon}_n$ depends on location $\hat{\varepsilon}_n := \hat{\varepsilon}_n(x)$. The fact that $\hat{\varepsilon}_n(x)$ converges to zero uniformly over $x \in D$ as $n \to \infty$ is enough to make the proof in [26] carry through with essentially no modifications. We remark that the arguments in [26] were also used in [14]; in [14] the presence of a nonconstant density makes computations a bit more tedious.

Step 3. Suppose that $u \in BV(D)$ is of the form $u = \mathbb{1}_A$ for some measurable set $A \subseteq D$. That is, suppose that u is the indicator function of a set with finite perimeter (with respect to D). In the appendix we show the following fact: there exists a sequence $\{\mathcal{F}_l\}_{l\in\mathbb{N}}$ of collections of closed balls contained in D satisfying the following:

1. For every $l \in \mathbb{N}$, the family $\mathcal{F}_l$ is finite.
2. For every $l \in \mathbb{N}$, the balls $B := \overline{B(x_B, r_B)}$ in $\mathcal{F}_l$ are pairwise disjoint.
3. $\lim_{l\to\infty} \max\{r_B \ : \ B \in \mathcal{F}_l\} = 0$.
4. For every $l \in \mathbb{N}$, and for every $B \in \mathcal{F}_l$, $|Du|(\partial B) = 0$.
5. $\lim_{l\to\infty} |Du|\big(\bigcup_{B\in\mathcal{F}_l} B\big) = |Du|(D)$.

To intuitively describe the properties of the family $\mathcal{F}_l$ suppose that A has smooth boundary ∂A. Then, for large l, the collection of balls $\mathcal{F}_l$ is a finite collection of disjoint small balls in $\mathbb{R}^d$ which essentially cover ∂A, and are such that $\partial A \cap \partial \bigcup_{B\in\mathcal{F}_l} B$ has zero surface area (in other words the surface of the balls and the surface of A are not aligned). For A with smooth boundary it is straightforward to see that such construction is possible, but there are some technical details one must take care of to extend the construction to more general sets. These details are presented in the appendix.

With the families $\{\mathcal{F}_l\}_{l\in\mathbb{N}}$ at hand, we can localize the problem and consider the functions

$$g_l(x) := \begin{cases} \min_{y\in B} g(y) & \text{if } x \in B, \quad B \in \mathcal{F}_l, \\ 0 \quad \text{otherwise} \end{cases}$$

for which we can see that

$$\begin{aligned}\liminf_{n\to\infty} E_n(u_n) &\geq \liminf_{n\to\infty} \sum_{B\in\mathcal{F}_l} \int_B \left(\frac{1}{\hat{\varepsilon}_n(x)} \int_B \eta_{\hat{\varepsilon}_n(x)}(|x-y|)|u_n(x)-u_n(y)|dy \right) g(x)dx \\ &\geq \sum_{B\in\mathcal{F}_l} \liminf_{n\to\infty} \int_B \left(\frac{1}{\hat{\varepsilon}_n(x)} \int_B \eta_{\hat{\varepsilon}_n(x)}(|x-y|)|u_n(x)-u_n(y)|dy \right) g(x)dx \\ &\geq \sum_{B\in\mathcal{F}_l} g_l(x_B) \liminf_{n\to\infty} \int_B \left(\frac{1}{\hat{\varepsilon}_n(x)} \int_B \eta_{\hat{\varepsilon}_n(x)}(|x-y|)|u_n(x)-u_n(y)|dy \right) dx \\ &\geq \sum_{B\in\mathcal{F}_l} g_l(x_B)\sigma_\eta TV_B(u) \\ &= \sum_{B\in\mathcal{F}_l} g_l(x_B)\sigma_\eta \int_B d|Du|(x) \\ &= \sigma_\eta \int_D g_l(x)d|Du|(x),\end{aligned}$$

where in the fourth inequality we used Step 2 and in the first equality we used Lemma 15.12 in [24] combined with property 4 in Step 3 for the balls in $\mathcal{F}_l$.

Finally, from properties 3 and 5 in Step 3 of $\{\mathcal{F}_l\}_{l\in\mathbb{N}}$ and from the Lipschitz continuity of g we know that

$$\lim_{l\to\infty} g_l(x) = g(x) \quad \text{for } |Du|\text{-a.e. } x\in D.$$

The dominated convergence theorem implies that

$$\lim_{l\to\infty} \int_D g_l(x)d|Du|(x) = \int_D g(x)d|Du|(x) = TV\,(u;g)\,,$$

where in the last equality we have used the representation formula (2.13). We conclude that for functions $u = \mathbb{1}_A \in BV(D)$ (i.e., indicator functions for sets of finite perimeter) inequality (3.1) holds.

Step 4. In order to prove (3.1) for general $u\in BV(D)$ we use Step 3 and the coarea formula for the energies E_n and F. Notice that the coarea formula for F is given in (2.14), whereas the coarea formula for E_n follows directly from the identity

$$|a-b| = \int_{-\infty}^{\infty} |\mathbb{1}_{a>t} - \mathbb{1}_{b>t}|dt \quad \forall a,b\in\mathbb{R}.$$

If $u_n \overset{L^1(D)}{\longrightarrow} u$, then for Lebesgue-a.e. $t\in\mathbb{R}$ we have

$$\mathbb{1}_{\{u_n>t\}} \overset{L^1(D)}{\longrightarrow} \mathbb{1}_{\{u>t\}} \quad \text{as } n\to\infty.$$

We may then apply Step 3 to conclude that for almost every $t\in\mathbb{R}$,

$$\liminf_{n\to\infty} E_n(\mathbb{1}_{\{u_n>t\}}) \geq F(\mathbb{1}_{\{u>t\}}).$$

Integrating the above inequality with respect to t, and using Fatou's lemma and the coarea formulas for the energies E_n and F, we deduce that

$$\begin{aligned}\liminf_{n\to\infty} E_n(u_n) = \liminf_{n\to\infty}\int_{-\infty}^{\infty} E_n(\mathbb{1}_{\{u_n>t\}})dt &\geq \int_{-\infty}^{\infty}\liminf_{n\to\infty} E_n(\mathbb{1}_{\{u_n>t\}})dt\\ &\geq \int_{-\infty}^{\infty} F(\mathbb{1}_{\{u>t\}})dt = F(u).\end{aligned}$$

Having established the liminf inequality we now proceed to establishing the limsup inequality. From Remark 2.20 and the density of $C_c^\infty(\mathbb{R}^d)$ functions in $L^1(D)$ with respect to TV (see Proposition 2.4 in [14]), it is enough to find a recovery sequence for every u obtained as the restriction to D of a function in $C_c^\infty(\mathbb{R}^d)$. We actually show that for every $u \in C_c^\infty(\mathbb{R}^d)$

$$\limsup_{n\to\infty} F_n(u) \leq F(u). \tag{3.3}$$

In other words, $u_n := u$ for all n is a recovery sequence for u. To prove this, notice that

$$\begin{aligned}E_n(u) &= \int_D \left(\frac{1}{\hat{\varepsilon}_n(x)}\int_{B(x,\hat{\varepsilon}_n(x))\cap D} \eta_{\hat{\varepsilon}_n(x)}(x-y)|u(x)-u(y)|dy\right) g(x)dx\\ &\leq \int_D \left(\frac{1}{\hat{\varepsilon}_n(x)}\int_{B(x,\hat{\varepsilon}_n(x))} \eta_{\hat{\varepsilon}_n(x)}(x-y)|u(x)-u(y)|dy\right) g(x)dx\\ &\leq \int_0^1\int_D \left(\frac{1}{\hat{\varepsilon}_n(x)}\int_{B(x,\hat{\varepsilon})} \eta_{\hat{\varepsilon}_n(x)}(x-y)|\nabla u(x+t(y-x))\cdot(x-y)|dy\right) g(x)dxdt,\end{aligned}$$

which follows from the fundamental theorem of calculus. Now, for every fixed $x \in D$, we use the change of variables $h := \frac{y-x}{\hat{\varepsilon}_n(x)}$ in the above expression to rewrite

$$\frac{1}{\hat{\varepsilon}_n(x)}\int_{B(x,\hat{\varepsilon}(x))} \eta_{\hat{\varepsilon}_n(x)}(x-y)|\nabla u(x+t(y-x))\cdot(x-y)|dy = \int_{B(0,1)} \eta(h)|\nabla u(x+t\hat{\varepsilon}_n(x)h)\cdot h|dh.$$

But because $u \in C_c^\infty(\mathbb{R}^d)$ and $\sup_{x\in D}\hat{\varepsilon}_n(x) \to 0$ as $n \to \infty$, it follows that

$$\lim_{n\to\infty}\sup_{x\in D}\left|\int_{B(0,1)} \eta(h)|\nabla u(x+t\hat{\varepsilon}_n(x)h)\cdot h|dh - \int_{B(0,1)}\eta(h)|\nabla u(x)\cdot h|dh\right| = 0.$$

Hence,

$$\limsup_{n\to\infty} E_n(u) \leq \int_0^1\int_D\left(\int_{B(0,1)}\eta(h)|\nabla u(x)\cdot h|dh\right)g(x)dxdt = \sigma_\eta TV(u;g).$$

This completes the proof. ■

With the previous lemma at hand we can now establish Proposition 3.1.

Proof of Proposition 3.1. We make use of Lemma 3.2 and for that purpose we first need to approximate the function $\rho : D \to \mathbb{R}$ from above and below using appropriate Lispchitz functions. More precisely, for every $s \in \mathbb{N}$ let

$$\tilde{\rho}_{1,s}(x) := \inf_{y \in D} \{\rho(y) + s|x-y|\}\,, \quad \rho_{2,s}(x) := \sup_{y \in D} \{\rho(y) - s|x-y|\}\,.$$

Notice that the functions $\tilde{\rho}_{1,s}, \rho_{2,s}$ are Lipschitz continuous, and as $s \to \infty$,

$$\rho_{1,s}(x) \nearrow \rho(x), \quad \rho_{2,s}(x) \searrow \rho(x) \quad \forall x \in D.$$

For every $s \in \mathbb{N}$ we modify the function $\tilde{\rho}_{1,s}$ slightly so as to create a Lipschitz function $\rho_{1,s}$ that in addition to minorizing ρ, is such that for all large enough $n \in \mathbb{N}$,

$$\int_{B(x,\varepsilon_n(x))} \rho_{1,s}(y)dy \leq \int_{B(x,\varepsilon_n(x)) \cap D} \rho(y)dy = \nu(B(x,\varepsilon_n(x))) \quad \forall x \in D. \tag{3.4}$$

To achieve this, we use the regularity assumption on the boundary of D as follows. From the fact that D is an open and bounded set with Lipschitz boundary it follows (see [19, Theorem 1.2.2.2]), that there exists a cone $\mathcal{C} \subseteq \mathbb{R}^d$ with nonempty interior and vertex at the origin, a family of rotations $\{R_x\}_{x \in D}$, and a number $1 > \zeta > 0$ such that for every $x \in D$,

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

For every $s \in \mathbb{N}$, let $\xi_s : \mathbb{R}^d \to [0,1]$ be a smooth cutoff function satisfying $\xi_s(x) \equiv 1$ in $\{y \in D \,:\, \mathrm{dist}(y,\partial D) > 1/s\}$ and $\xi \equiv 0$ in $\{y \in D \,:\, \mathrm{dist}(y,\partial D) < 1/2s\}$, and let us define

$$\rho_{1,s}(x) := \frac{|B(0,1) \cap \mathcal{C}|}{|B(0,1)|}\rho_{min}(1-\xi_s(x)) + \tilde{\rho}_{1,s}(x)\xi_s(x), \quad x \in \mathbb{R}^d.$$

It is clear that the functions $\rho_{1,s}$ are Lispchitz, they minorize ρ, and they converge to ρ pointwise. In addition, for any fixed s, if $n \in \mathbb{N}$ is large enough so that, in particular, $\sup_{x \in D} \varepsilon_n(x) < \min\{\zeta, \frac{1}{8s}\}$, then (3.4) holds. Indeed, notice that if $x \in D$ is $1/4s$ units away from ∂D, then (3.4) follows directly from the fact that $\rho_{1,s} \leq \rho$; on the other hand, if $x \in D$ is within distance $1/4s$ from ∂D, we conclude that

$$\begin{aligned}\int_{B(x,\hat{\varepsilon}_n(x))} \rho_{1,s}(y)dy = |\mathcal{C} \cap B(0,1)|\rho_{min}(\hat{\varepsilon}_n(x))^d &\leq \int_{(x+R_x(\mathcal{C})) \cap B(x,\hat{\varepsilon}_n(x))} \rho(y)dy \\ &\leq \int_{B(x,\hat{\varepsilon}_n(x)) \cap D} \rho(y)dy = \nu(B(x,\hat{\varepsilon}_n(x))).\end{aligned}$$

Let us now establish the limsup inequality. We introduce the functionals $F_{n,s} : L^1(D) \to [0,\infty)$ given by

$$F_{n,s}(u) := \int_D \tilde{f}_{n,s}(x) \left(\int_D \eta_{\varepsilon_n(x)}(x-y)|u(x)-u(y)|\rho_{2,s}(y)dy \right) \rho_{2,s}(x)dx, \quad u \in L^1(D),$$

where

$$\tilde{f}_{n,s}(x) := \frac{(\tilde{\varepsilon}_n(x))^d}{(\mu_s(B(x,\varepsilon_n(x))))^{1+1/d}}$$

and

$$d\mu_s(x) := \rho_{1,s}(x)dx.$$

It follows that for every fixed $s \in \mathbb{N}$ for all large $n \in \mathbb{N}$,

$$F_n(v) \leq F_{n,s}(v) \quad \forall v \in L^1(D).$$

Hence, from (3.3) it follows that for every arbitrary $u \in C_c^\infty(\mathbb{R}^d)$,

$$\limsup_{n\to\infty} F_n(u) \leq \limsup_{n\to\infty} F_{n,s}(u) \leq \frac{\sigma_\eta}{\alpha_d^{1+1/d}} \int_D \frac{\rho_{2,s}^2(x)}{\rho_{1,s}^{1+1/d}(x)} d|Du|(x).$$

Taking $s \to \infty$ in the above inequality we deduce that

$$\limsup_{n\to\infty} F_n(u) \leq \frac{\sigma_\eta}{\alpha_d^{1+1/d}} TV(u;\rho^{1-1/d}).$$

Remark 2.20 and the density of $C_c^\infty(\mathbb{R}^d)$ functions with respect to TV imply the desired result.

To establish the liminf inequality, it is enough to change the roles of $\rho_{1,s}$ and $\rho_{2,s}$ in the definition of the functional $F_{n,s}$ and use Lemma 3.2 to conclude that if $u_n \overset{L^1(D)}{\longrightarrow} u$, then for every $s \in \mathbb{N}$

$$\liminf_{n\to\infty} F_n(u_n) \geq \liminf_{n\to\infty} F_{n,s}(u_n) \geq \frac{\sigma_\eta}{\alpha_d^{1+1/d}} \int_D \frac{\rho_{1,s}^2(x)}{\rho_{2,s}^{1+1/d}(x)} d|Du|(x).$$

Taking the limit as $s \to \infty$ on the right-hand side of the above expression we obtain the desired result. ■

Let us pause for a moment and outline the general plan for the remainder of the proof. We will establish two inequalities of the form

$$GTV_{n,k_n}(u_n) \leq (1+o(n))F_n(u_n \circ T_n)$$

and

$$F_n(u_n \circ T_n) \leq (1+o(n))GTV_{n,k_n}(u_n),$$

where T_n is the transport map with controlled ∞-OT (optimal transport) cost in (2.11), and F_n is the nonlocal energy introduced in Proposition 3.1 for appropriately picked functions $\varepsilon_n(\cdot)$ and $\hat{\varepsilon}_n(\cdot)$; the function $\varepsilon_n(x)$ is chosen so as to guarantee that the number of sample points in the ball $B(x,\varepsilon_n(x))$ is k_n, and $\hat{\varepsilon}_n$ is asymptotically uniformly equivalent to ε_n. Indeed, since ε_n is chosen to be much larger than $\delta_n = \|Id - T_n\|_\infty$, we will be able to expand or contract the averaging region in the nonlocal energies at our convenience in order to enforce the desired inequalities. The bottom line is that if $u_n \to_{TL^1} u$, then $u_n \circ T_n \to_{L^1} u$, and thus we can combine the above inequalities with Proposition 3.1 to obtain the liminf and limsup inequalities needed to establish Theorem 2.6.

Proof of liminf inequality in Theorem 2.6. Let us consider the cone $\mathcal{C}$, the family of rotations $\{R_x\}_{x\in D}$, and the number $1 > \zeta > 0$ introduced in the proof of Proposition 3.1. For $0 \leq \gamma_1 < \gamma_2 < \zeta$ and $x \in D$, let $A(x;\gamma_1,\gamma_2)$ be the annulus $A(x;\gamma_1,\gamma_2) := B(x,\gamma_2) \setminus B(x,\gamma_1)$. Then, there exist constants $\mathcal{K}_1$ (only depending on d, and on the cone $\mathcal{C}$) and $\mathcal{K}_2$ (only depending on d), such that for every $0 \leq \gamma_1 < \gamma_2 < \zeta$,

$$\mathcal{K}_1\rho_{min}(\gamma_2)^{d-1}\cdot(\gamma_2-\gamma_1) \leq \nu\left(A(x;\gamma_1,\gamma_2)\cap(x+R_x(\mathcal{C}))\right) \leq \nu\left(A(x;\gamma_1,\gamma_2)\right) \quad \forall x\in D \tag{3.5}$$

and

$$\nu\left(A(x;\gamma_1,\gamma_2)\right) \leq \mathcal{K}_2\rho_{max}(\gamma_2)^{d-1}\cdot(\gamma_2-\gamma_1) \quad \forall x \in D. \tag{3.6}$$

We work in a set with full probability for which the maps $\{T_n\}_{n\in\mathbb{N}}$ from (2.11) exist. For every $x \in D$ we let $\varepsilon_n(x)$ be the number for which

$$\nu\left(B(x,\varepsilon_n(x))\right) = \bar{\varepsilon}_n^d = \frac{k_n}{n}.$$

On the other hand, we define $\dot{\varepsilon}_n(x)$ and $\hat{\varepsilon}_n(x)$ to be

$$\dot{\varepsilon}_n(x) := \varepsilon_n(x) - \frac{\mathcal{K}_2\rho_{max}\delta_n}{\mathcal{K}_1\rho_{min}},$$

$$\hat{\varepsilon}_n(x) := \dot{\varepsilon}_n(x) - 3\delta_n,$$

where δ_n is as in (2.11).

Notice that the assumption $n \gg k_n \gg \log(n)$ is equivalent to $1 \gg \bar{\varepsilon}_n \gg \delta_n$ where we recall $\bar{\varepsilon}_n$ is defined in (2.3). Combining this with (3.5) we deduce that

$$\lim_{n\to\infty}\sup_{x\in D}\varepsilon_n(x) = 0, \quad \lim_{n\to\infty}\sup_{x\in D}\left|\frac{\varepsilon_n(x)}{\hat{\varepsilon}_n(x)} - 1\right| = 0, \quad \lim_{n\to\infty}\frac{\delta_n}{\inf_{x\in D}\hat{\varepsilon}_n(x)} = 0. \tag{3.7}$$

Now, from (3.5) we see that for all large enough n, and for all $x \in D$,

$$\begin{aligned}\nu(B(x,\dot{\varepsilon}_n(x))) &= \nu(B(x,\varepsilon_n(x))) - \nu(A(x;\dot{\varepsilon}_n(x),\varepsilon_n(x)))\\ &\leq \frac{k_n}{n} - \mathcal{K}_2\rho_{max}(\varepsilon_n(x))^{d-1}\delta_n.\end{aligned} \tag{3.8}$$

Thus,

$$\begin{aligned}\nu_n(B(x,\dot{\varepsilon}_n(x))) &= \nu_n(B(x,\dot{\varepsilon}_n(x))) - \nu(B(x,\dot{\varepsilon}_n(x))) + \nu(B(x,\dot{\varepsilon}_n(x)))\\ &\leq \nu(B(x,\dot{\varepsilon}_n(x) + \|Id - T_n\|_\infty)) - \nu(B(x,\dot{\varepsilon}_n(x))) + \nu(B(x,\dot{\varepsilon}_n(x)))\\ &\leq \mathcal{K}_2\rho_{max}\varepsilon_n(x)^{d-1}\|Id-T_n\|_\infty + \frac{k_n}{n} - \mathcal{K}_2\rho_{max}(\varepsilon_n(x))^{d-1}\cdot\delta_n\\ &\leq \mathcal{K}_2\rho_{max}\varepsilon_n(x)^{d-1}\delta_n + \frac{k_n}{n} - \mathcal{K}_2\rho_{max}\varepsilon_n(x)^{d-1}\delta_n\\ &= \frac{k_n}{n},\end{aligned}$$

where in the first inequality we have used the fact that T_n is a transportation map between ν and ν_n, and in the second inequality we have used (3.6). Combining the previous inequality with the fact that $B(T_n(x), \dot{\varepsilon}_n(x) - \delta_n) \subseteq B(x, \dot{\varepsilon}_n(x))$ we deduce that

$$\nu_n(B(T_n(x), \hat{\varepsilon}_n(x) - \delta_n)) \leq \frac{k_n}{n}.$$

In particular, if $y \in D$ is such that $T_n(y) \not\sim_{k_n} T_n(x)$, then from Remark 2.1 it follows that

$$\hat{\varepsilon}_n(x) = \dot{\varepsilon}_n(x) - 3\delta_n < |T_n(x) - T_n(y)| - 2\delta_n \leq |T_n(x) - T_n(y)| - |x - T_n(x)| - |y - T_n(y)| \leq |x - y|.$$

We conclude that for all large enough n, $J_{k_n}(T_n(x), T_n(y)) = 0$ implies $\hat{\varepsilon}_n(x) < |x - y|$. In particular, we deduce that for all large enough $n \in \mathbb{N}$,

$$J_{k_n}(T_n(x), T_n(y)) \geq \eta\left(\frac{|x-y|}{\hat{\varepsilon}_n(x)}\right) \quad \forall x, y \in D. \tag{3.9}$$

Summarizing, by contracting the radius $\varepsilon_n(x)$ a little bit we obtained $\hat{\varepsilon}_n(x)$ satisfying (3.7) and (3.9).

Let us now consider a sequence of functions $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in L^1(\nu_n)$ for which $u_n \xrightarrow{TL^1} u$ for some $u \in L^1(\nu)$. It follows from (3.9) that for all large enough $n \in \mathbb{N}$,

$$\begin{aligned}
GTV_{n,k_n}(u_n) &= \frac{1}{n^2 \bar{\varepsilon}_n^{d+1}} \sum_{i,j} J_{k_n}(\mathbf{x}_i, \mathbf{x}_j)|u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)| \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{\bar{\varepsilon}_n^{d+1}} \left(\sum_{j=1}^{n} J_{k_n}(\mathbf{x}_i, \mathbf{x}_j)|u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)| \right) \\
&= \int_D \frac{1}{\bar{\varepsilon}_n^{d+1}} \left(\int_D J_{k_n}(T_n(x), T_n(y))|u_n \circ T_n(x) - u_n \circ T_n(y)| d\nu(y) \right) d\nu(x) \\
&\geq \int_D \frac{1}{\bar{\varepsilon}_n^{d+1}} \left(\int_D \eta\left(\frac{|x-y|}{\hat{\varepsilon}_n(x)}\right) |u_n \circ T_n(x) - u_n \circ T_n(y)| d\nu(y) \right) d\nu(x) \\
&= \int_D \hat{f}_n(x) \left(\int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|) \, |u_n \circ T_n(x) - u_n \circ T_n(y)| d\nu(y) \right) d\nu(x) \\
&= F_n(u_n \circ T_n),
\end{aligned} \tag{3.10}$$

where the function $\hat{f}_n$ is given by $\hat{f}_n(x) = \frac{\hat{\varepsilon}_n(x)^d}{(\nu(B(x,\varepsilon_n(x))))^{1+1/d}}$, and for an arbitrary function $v \in L^1(\nu)$, $F_n(v)$ is defined as

$$F_n(v) := \int_D \hat{f}_n(x) \left(\int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|) \, |v(x) - v(y)| d\nu(y) \right) d\nu(x).$$

The last equality follows after recalling that $\varepsilon_n(x)$ was defined to satisfy $\nu(B(x, \varepsilon_n(x))) = \bar{\varepsilon}_n$.

Since $u_n \xrightarrow{TL^1} u$ implies that $u_n \circ T_n \xrightarrow{L^1(\nu)} u$, it follows from Proposition 3.1 and from (3.10) that

$$\liminf_{n\to\infty} GTV_{n,k_n}(u_n) \geq \liminf_{n\to\infty} F_n(u_n \circ T_n) \geq \frac{\sigma_\eta}{\alpha_d^{1+1/d}} TV(u; \rho^{1-1/d}),$$

which establishes the liminf inequality in Theorem 2.6. ■

Proof of the limsup *inequality in Theorem* 2.6. We work in a set with full probability for which the maps $\{T_n\}_{n\in\mathbb{N}}$ from (2.11) exist. By a construction analogue to the one used in the proof of the liminf inequality (this time enlarging $\varepsilon_n(x)$ instead of contracting it), we may construct a function $\hat{\varepsilon}_n$ satisfying (3.7) and

$$J_{k_n}(T_n(x), T_n(y)) \leq \eta\left(\frac{|x-y|}{\hat{\varepsilon}_n(x)}\right) \quad \forall x, y \in D.$$

Let $u \in C_c^\infty(\mathbb{R}^d)$ and for every $n \in \mathbb{N}$ let u_n be the function in $L^1(\nu_n)$ defined by

$$u_n(\mathbf{x}_i) := u(\mathbf{x}_i), \quad i = 1, \ldots, n.$$

In other words, u_n is the function u restricted to the point cloud X_n. Then,

$$\begin{aligned}
GTV_{n,k_n}(u_n) &= \frac{1}{n^2\bar{\varepsilon}_n^{d+1}} \sum_{i,j} J(\mathbf{x}_i, \mathbf{x}_j)|u(\mathbf{x}_i) - u(\mathbf{x}_j)| \\
&= \int_D \frac{1}{\bar{\varepsilon}_n^{d+1}} \left(\int_D J(T_n(x), T_n(y))|u \circ T_n(x) - u \circ T_n(y)|d\nu(y)\right) d\nu(x) \\
&\leq \int_D \frac{1}{\bar{\varepsilon}_n^{d+1}} \left(\int_D \eta\left(\frac{|x-y|}{\hat{\varepsilon}_n(x)}\right) |u \circ T_n(x) - u \circ T_n(y)|d\nu(y)\right) d\nu(x) \qquad (3.11)\\
&\leq \int_D \hat{f}_n(x) \left(\int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|)\,|u(x) - u(y)|d\nu(y)\right) d\nu(x) + \mathcal{I}_n + \mathcal{II}_n \\
&= F_n(u) + \mathcal{I}_n + \mathcal{II}_n,
\end{aligned}$$

where

$$\mathcal{I}_n := \int_D \hat{f}_n(x) \left(\int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|)\,|u(x) - u(T_n(x))|d\nu(y)\right) d\nu(x)$$

and

$$\mathcal{II}_n := \int_D \hat{f}_n(x) \left(\int_D \eta_{\hat{\varepsilon}_n(x)}(|x-y|)\,|u(y) - u(T_n(y))|d\nu(y)\right) d\nu(x),$$

but notice that

$$\mathcal{I}_n, \mathcal{II}_n \leq \frac{C\|\nabla u\|_\infty \|Id - T_n\|_\infty}{\inf_{x\in D}\hat{\varepsilon}_n(x)}$$

for some constant C. Hence, $\lim_{n\to\infty}\mathcal{I}_n = \lim_{n\to\infty}\mathcal{II}_n = 0$. It follows from the proof of Proposition 3.1 that

$$\limsup_{n\to\infty} GTV_{n,k_n}(u_n) \leq \limsup_{n\to\infty} F_n(u) \leq \frac{\sigma_\eta}{\alpha_d^{1+1/d}} TV(u; \rho^{1-1/d}).$$

Using the density of $C_c^\infty(\mathbb{R}^d)$ with respect to TV, we may find a recovery sequence $\{u_n\}_{n\in\mathbb{N}}$ with $u_n \in L^1(\nu_n)$ for every $u \in L^1(D)$. Finally, if $u \in L^1(D)$ is of the form $u = \mathbb{1}_A$ for some measurable set $A \subseteq D$, we can actually choose this recovery sequence to consist of indicator functions, as it follows from the fact that the energies GTV_{n,k_n} satisfy a coarea formula (see the proof of Corollary 1.3 in [14]). This establishes the last statement in Theorem 2.6. ■

3.1. Proof of compactness.

Proof of Theorem 2.7. The compactness in Theorem 2.7 follows directly from the compactness result in [14]. Indeed, consider $\hat{\varepsilon}_n$ as defined in the proof of the liminf inequality of Theorem 2.6 and define

$$\varepsilon_n^l := \inf_{x \in D} \hat{\varepsilon}_n(x) - 2\delta_n.$$

Then it follows that for all large enough n,

$$\frac{\varepsilon_n^l}{\overline{\varepsilon}_n} \geq c > 0$$

for some constant $c > 0$. Moreover, for every i, j we have

$$J_{k_n}(\mathbf{x}_i, \mathbf{x}_j) \geq \eta\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{\varepsilon_n^l}\right). \tag{3.12}$$

Let $\{u_n\}_{n\in\mathbb{N}}$ be a sequence with $u_n \in L^1(\nu_n)$ satisfying

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

and

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

It follows from inequality (3.12) that

$$GTV_{n,k_n}(u_n) \geq \frac{c^d}{n^2 \varepsilon_n^l} \sum_{i,j} \eta_{\varepsilon_n^l}(|\mathbf{x}_i - \mathbf{x}_j|)|u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)| =: c^d GTV_{n,\varepsilon_n^l}(u_n).$$

Therefore,

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

The precompactness of $\{u_n\}_{n\in\mathbb{N}}$ in TL^1 follows from Theorem 1.2 in [14]. ■

Appendix A. In this appendix we prove the claim we made in Step 3 in the proof of Lemma 3.2. Let us start by noting that because A has finite perimeter, we may talk about its *reduced boundary* $\partial^* A$ (see [24]). In particular, for every measurable set $S \subseteq D$,

$$|Du|(S) = \mathcal{H}^{d-1}(\partial^* A \cap S),$$

where $\mathcal{H}^{d-1}$ stands for the $(d-1)$-dimensional Hausdorff measure.

De Giorgi's structure theorem (see Theorem 15.9 in [24]) implies that there exist countably many C^1-hypersurfaces $\{M_i\}_{i\in\mathbb{N}}$, compact sets $K_i \subseteq M_i$, and a $\mathcal{H}^{d-1}$-null set F such that

$$\partial^* A = F \cup \bigcup_{i\in\mathbb{N}} K_i.$$

Fix $j \in \mathbb{N}$. From the fact that the sets K_i are subsets of C^1-hypersurfaces, it is straightforward to see that one can find a sequence $\{\tilde{\mathcal{F}}_l^j\}_{l \in \mathbb{N}}$ of families of closed balls contained in D, satisfying properties 1, 2, and 3 in Step 3 as above and such that

$$\lim_{l \to \infty} \mathcal{H}^{d-1}\left(\bigcup_{B \in \tilde{\mathcal{F}}_l^j} B \cap \bigcup_{i=1}^{j} K_i \right) = \mathcal{H}^{d-1}\left(D \cap \bigcup_{i=1}^{j} K_i \right).$$

Still working with j fixed, we now proceed to modify the balls in the families $\tilde{\mathcal{F}}_l^j$ to obtain balls that also satisfy property 4 in Step 3. Indeed, we consider the family $\mathcal{F}_l^j$ of closed balls obtained from the balls in $\tilde{\mathcal{F}}_l^j$ by adding δ_l units to their original radius. The number δ_l can be chosen in such a way that $0 < \delta_l < \frac{1}{l}$, and the balls in $\mathcal{F}_l^j$ are all contained in D, are pairwise disjoint, and satisfy property 4 in Step 3. This can be done because $\partial^* A \cap D$ has finite $\mathcal{H}^{d-1}$-measure and hence its intersection with the boundary of arbitrary balls centered at a fixed point can only have nonzero $\mathcal{H}^{d-1}$-measure for at most countably many of such balls.

A sequence of families $\{\mathcal{F}_l\}_{l \in \mathbb{N}}$ of balls with the desired properties can then be obtained from the families $\{\mathcal{F}_l^j\}_{l \in \mathbb{N}}$ for every $j \in \mathbb{N}$, by using a diagonal argument knowing that

$$\lim_{j \to \infty} \mathcal{H}^{d-1}\left(D \cap \bigcup_{i=1}^{j} K_i \right) = \mathcal{H}^{d-1}\left(D \cap \partial^* A \right).$$

REFERENCES

[1] A. E. Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan, *Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning*, in 29th Annual Conference on Learning Theory, V. Feldman, A. Rakhlin, and O. Shamir, eds., Proceedings of Machine Learning Research 49, Columbia University, New York, 2016, pp. 879–906.

[2] E. Arias-Castro, B. Pelletier, and P. Pudlo, *The normalized graph cut and Cheeger constant: From discrete to continuous*, Adv. in Appl. Probab., 44 (2012), pp. 907–937.

[3] A. Baldi, *Weighted BV functions*, Houston J. Math., 27 (2001), pp. 683–705.

[4] M. Belkin and P. Niyogi, *Convergence of Laplacian eigenmaps*, in Advances in Neural Information Processing Systems (NIPS) 19, MIT Press, Cambridge, MA, 2007, pp. 129–136.

[5] M. Belkin and P. Niyogi, *Towards a theoretical foundation for Laplacian-based manifold methods*, J. Comput. System Sci., 74 (2008), pp. 1289–1308.

[6] A. L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis, *Uncertainty Quantification in the Classification of High Dimensional Data*, preprint, https://arxiv.org/abs/1703.08816, 2017.

[7] X. Bresson, T. Laurent, D. Uminsky, and J. von Brecht, *Multiclass total variation clustering*, in Advances in Neural Information Processing Systems 26, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds., 2013, pp. 1421–1429.

[8] J. Calder, *The Game Theoretic p-Laplacian and Semi-Supervised Learning with Few Labels*, preprint, https://arxiv.org/abs/1711.10144, 2017.

[9] R. R. Coifman and S. Lafon, *Diffusion maps*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 5–30.

[10] G. Dal Maso, *An Introduction to Γ-Convergence*, Birkhäuser, Basel, 1993.

[11] N. García Trillos, Z. Kaplan, T. Samkhoana, and D. Sanz-Alonso, *On the Consistency of Graph-Based Bayesian Learning and the Scalability of Sampling Algorithms*, preprint, https://arxiv.org/abs/1710.07702, 2017.

[12] N. García Trillos and R. Murray, *A New Analytical Approach to Consistency and Overfitting in Regularized Empirical Risk Minimization*, preprint, https://arxiv.org/abs/1607.00274, 2016.

[13] N. García Trillos and D. Sanz-Alonso, *Continuum Limit of Posteriors in Graph Bayesian Inverse Problems*, preprint, https://arxiv.org/abs/1706.07193, 2017.

[14] N. García Trillos and D. Slepčev, *Continuum limit of total variation on point clouds*, Arch. Ration. Mech. Anal., 220 (2016), pp. 193–241.

[15] N. García Trillos and D. Slepčev, *On the rate of convergence of empirical measures in* ∞*-transportation distance*, Canad. J. Math., 67 (2015), pp. 1358–1383.

[16] N. García Trillos and D. Slepčev, *A Variational Approach to the Consistency of Spectral Clustering*, preprint, https://arxiv.org/abs/1508.01928, 2015.

[17] N. García Trillos, D. Slepčev, J. von Brecht, T. Laurent, and X. Bresson, *Consistency of Cheeger and ratio graph cuts*, J. Mach. Learn. Res., 17 (2016), 181.

[18] E. Giné and V. Koltchinskii, *Empirical graph Laplacian approximation of Laplace-Beltrami operators: Large sample results*, in High Dimensional Probability, IMS Lecture Notes Monogr. Ser. 51, Inst. Math. Statist., Beachwood, OH, 2006, pp. 238–259.

[19] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman (Advanced Publishing Program), Boston, 1985.

[20] M. Hein, J.-Y. Audibert, and U. von Luxburg, *From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians*, in Learning Theory, Lecture Notes in Comput. Sci. 3559, Springer, Berlin, 2005, pp. 470–485.

[21] 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), 2010, pp. 847–855.

[22] M. Hein and S. Setzer, *Beyond spectral clustering—tight relaxations of balanced graph cuts*, in Proccedings of Advances in Neural Information Processing Systems (NIPS), J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds., 2011, pp. 2366–2374.

[23] G. Leoni, *A First Course in Sobolev Spaces*, Grad. Stud. Math. 105, American Mathematical Society, Providence, RI, 2009.

[24] F. Maggi, *Sets of Finite Perimeter and Geometric Variational Problems*, Cambridge Stud. Adv. Math. 135, Cambridge University Press, Cambridge, 2012.

[25] M. Penrose, *Random Geometric Graphs*, Oxford Stud. Probab. 5, Oxford University Press, Oxford, 2003.

[26] A. C. Ponce, *A new approach to Sobolev spaces and connections to* Γ*-convergence*, Calc. Var. Partial Differential Equations, 19 (2004), pp. 229–255.

[27] S. S. Rangapuram and M. Hein, *Constrained 1-spectral clustering*, in Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 2012, pp. 1143–1151.

[28] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), pp. 888–905.

[29] A. Singer, *From graph to manifold Laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21 (2006), pp. 128–134.

[30] A. Singer and H.-T. Wu, *Spectral convergence of the connection Laplacian from random samples*, Inf. Inference, 6 (2017), pp. 58–123.

[31] D. Slepčev and M. Thorpe, *Analysis of p-Laplacian Regularization in Semi-Supervised Learning*, preprint, https://arxiv.org/abs/1707.06213, 2017.

[32] 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, pp. 1079–1086.

[33] U. von Luxburg, *A tutorial on spectral clustering*, Stat. Comput., 17 (2007), pp. 395–416.

[34] U. von Luxburg, M. Belkin, and O. Bousquet, *Consistency of spectral clustering*, Ann. Statist., 36 (2008), pp. 555–586.