# Asymptotic analysis of the Ginzburg–Landau functional on point clouds

## Matthew Thorpe[*] and Florian Theil

University of Warwick, Coventry, CV4 7AL, UK
(m.thorpe@maths.cam.ac.uk)

The Ginzburg–Landau functional is a phase transition model which is suitable for classification type problems. We study the asymptotics of a sequence of Ginzburg–Landau functionals with anisotropic interaction potentials on point clouds $\Psi_n$ where $n$ denotes the number data points. In particular, we show the limiting problem, in the sense of $\Gamma$-convergence, is related to the total variation norm restricted to functions taking binary values, which can be understood as a surface energy. We generalize the result known for isotropic interaction potentials to the anisotropic case and add a result concerning the rate of convergence.

## 1. Introduction

### 1.1. Finite dimensional modelling

In the age of 'big data' the mathematical modeller is often without a physical model and instead uses a data driven approach for which graphical models are a powerful tool. Graphical based modelling techniques are used across a very broad spectrum of problems from social science type problems, such as identifying communities [**18**, **22**, **39**, **43**, **46**], to image segmentation [**6**, **29**], to cell biology [**14**], to modelling the world wide web [**7**, **12**, **14**, **21**] and many more. Anisotropic models, studied in this paper, have found applications in cosmological models [**28**, **34**], modelling outbreaks of disease [**33**] and image recognition [**47**].

Graphical models are based upon pairwise similarities which practitioners can design based on expert knowledge. The measure of similarity (on pairs) then defines a geometry (on a data set). The motivation in this paper is to consider a graphical approach to the classification problems. Given a measure of similarity, we wish to define a labelling using the geometry of the graph.

The problem is given data $\Psi_n = \{\xi_i\}_{i=1}^n \subset X$, where $X \subset \mathbb{R}^d$, find $\mu : \Psi_n \to \mathbb{R}$ that labels each data point. The labelling is constructed so that $\mu(\xi_i) = 0$ means

---

[*]Present address: Department of Applied Mathematics and Theoretical Physics, University of Cambridge, email: m.thorpe@maths.cam.ac.uk

that $\xi_i$ is associated with the class 0 and $\mu(\xi_i) = 1$ means that $\xi_i$ is associated with the class 1. For a finite number of observations, we allow a soft classification, however, the scaling is chosen such that in the data-rich limit classifiers are binary valued. The motivation for our approach is to validate approximating the hard classification problem by a soft classification problem. The soft classification problem is, in general, numerically easier [24] and therefore, more appealing to the practitioner. However, one also wants to be precise in regards to which class a data point belongs. Minimizers of the Ginzburg–Landau functional are used as a classification tool [42] in order to allow for phase transitions which allow a soft classification approach whilst also penalizing states that are not close to a hard classification.

Another important application for this work is in designing classifiers. By not assuming that the model is isotropic, we allow greater flexibility which allows one to choose some features as more important than others. The next subsection contains a simple example which shows how the design choice can affect the classification.

Assessing the validity of such an approach is of high importance. This is especially true as there is no natural link between the data generating process and the choice of the classifier. In particular, we argue that although using the Ginzburg–Landau functional is a good choice due to its phase transition properties, it is by no means the only available option. When one can make a connection between classifier and data generating process, for example, maximum likelihood estimators, then this link motivates the methodology. Without such a connection one needs to do more, such as show the estimators have desirable properties, in order to justify the approach. Other approaches that use classifiers that are detached from the data generating process include [41] where the authors prove the convergence of the $k$-means method using similar variational techniques.

An important criterion for validating the model is the behaviour in the large data limit. When increasing the size of the data set one should expect to see stability in classifiers. In particular, this requires convergence in the large data limit and the existence of a limiting (data rich) model. When one has a data generating model, that is, there is some truth, then one can talk about consistency. In the situation considered in this paper, there is no truth so instead, we use solutions to the limiting model. Knowledge of the limiting model gives an insight into what features one should expect for estimates from the finite data problem. In particular, this paper considers three questions:

(P1)  Do estimated classifiers $\mu_n$ converge as $n \to \infty$?

(P2)  Can we attach some meaning to any limit of $\mu_n$? That is, does there exists a limiting problem?

(P3)  Can we characterize the rate of convergence of estimators?

The primary results of this paper concern the first two questions. It is shown that estimates $\mu_n$ converge to the solution of a limiting problem. Furthermore, solutions to the limiting problem are binary valued which means we expect estimates $\mu_n$ for large $n$ to be approximately hard classifiers. For the third question, we give some preliminary results into characterizing the rate in a simplified example. We believe

Figure 1. An example graph. For the classifier estimates see figure 2.

these results will hold under more generality than stated here and it is the objective of ongoing work to extend them.

Our approach is motivated by [**1**, **25**, **42**]. Classifiers are constructed as the solution of a variational problem which is common in statistical problems, for example, maximum likelihood and maximum-a-posterior problems. In particular, minimizers of the Ginzburg–Landau functional, a phase transition model popular in material science and image segmentation, are used as classifiers. In the context of the 2-class classification problem the two phases are the classes and the phase transition corresponds to the set $\{x \in \Psi_n : \mu_n(x) \in (c, 1 - c)\}$ for some $c \in (0, 1)$, for example, $c = 0.1$. This is the subset of $\Psi_n$ that are not strongly associated with either class.

Classifiers $\mu_n : \Psi_n \to \mathbb{R}$ are constructed as follows. Let $V : \mathbb{R} \to [0, \infty)$ be a potential such that states taking the value 0 or 1 is favoured. For example, $V(t) = t^2(t - 1)^2$. A graph is constructed by taking the vertices as the set $\Psi_n = \{\xi_i\}_{i=1}^n \subset \mathbb{R}^d$ and weighting edges

$$W_{ij} = \eta_\epsilon(\xi_i - \xi_j) := \frac{1}{\epsilon^d} \varphi \circ \pi \left( \frac{\xi_i - \xi_j}{\epsilon} \right) \tag{1}$$

where $\pi : X \to \mathbb{R}$ is a given one-dimensional map so that $\pi(\xi_i)$ represents a feature of $\xi_i$ and $\varphi : \mathbb{R} \to [0, \infty)$ penalizes the difference $\pi(\xi_i - \xi_j)$. We say that there is an edge between $\xi_i$ and $\xi_j$ if $W_{ij} > 0$, for example, see figure 1. For a function $\mu_n$ on $\Psi_n$ the graph energy $\mathcal{E}_n(\mu_n) \in [0, \infty]$ is defined by

$$\mathcal{E}_n(\mu_n) = \frac{1}{\epsilon_n} \frac{1}{n} \sum_{i=1}^n V(\mu_n(\xi_i)) + \frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{i,j} W_{ij} |\mu_n(\xi_i) - \mu_n(\xi_j)|. \tag{2}$$

Our classifier is given as the minimizer of $\mathcal{E}_n$.

We call the map $\pi : X \to \mathbb{R}$ the feature projection as it allows the practitioner to include feature selection of the data $\xi_i$. For example, one may decide that two data points should be considered similar based on the pairwise difference. In this case, an appropriate choice would be the weighted Euclidean distance $\pi(x) = \sqrt{\sum_{i=1}^d w_i |x_i|^2}$. The isotropic case would correspond to weights $w_i = 1$. Other choices could be to include correlations between dimensions, for example, $\pi(x) = \sqrt{\sum_{i,j=1}^d w_{ij} |x_i| |x_j|}$.

The authors of [**25**] study the asymptotic properties of the graph total variation defined by

$$GTV_n(\mu_n) := \frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{i,j} W_{ij} |\mu_n(\xi_i) - \mu_n(\xi_j)| \tag{3}$$

when $W_{ij}$ is isotropic, that is, $\pi(x) = |x|$. In the special case that $\mu_n(\xi_i) \in \{0, 1\}$ this reduces to the graph cut of $\Psi_n$, that is, if $(\mu_n)^{-1}(0) = A_0$ and $(\mu_n)^{-1}(1) = A_1$ then

$$GTV_n(\mu_n) = \frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{\xi_i \in A_0} \sum_{\xi_j \in A_1} W_{ij}.$$

In particular, the authors in [**25**] show the $\Gamma$-convergence of $GTV_n$ to a weighted total variation $TV(\cdot; \rho, \eta)$ given by

$$TV(\mu; \rho, \eta) := \sigma_\eta \sup \left\{ \int_X \mu \operatorname{div}(\phi) \, dx \,:\, |\phi(x)| \leqslant \rho^2(x) \, \forall x \in X, \phi \in C_c^\infty(X; \mathbb{R}^d) \right\},$$

and $L^1$-compactness for any sequence $\mu_n$ with $\sup_n(GTV_n(\mu_n) + \|\mu_n\|_{L^1}) < \infty$.

We wish to allow for soft classification and the total variation term alone is not enough to be able to do this informatively. The classification approach is made more robust by including a first order term $V : \mathbb{R} \to [0, \infty)$ which penalizes associating a data point to more than one class. See, for example, figure 2 for a comparison. It is not trivial that the convergence results in [**25**] will survive adding a penalty term.

Finding minimizers of $\mathcal{E}_n$ is also an important problem but is not addressed in this paper. We instead refer to [**9**, **10**] for numerical methods.

## 1.2. Example: classification dependence on the choice of $\eta$

Through a toy problem, we demonstrate how the interaction potential can be used to pick out features of the practitioners choice. Data are points $\xi_i = (\xi_{1i}, \xi_{2i}) \in \mathbb{R}^2$ generated from four classes. For a fixed $\alpha$ the feature projection $\pi : \mathbb{R}^2 \to [0, \infty)$ is defined by the weighted Euclidean norm

$$\pi(\xi) = \sqrt{(1 - \alpha)\xi_1^2 + \alpha\xi_2^2}.$$

For $\alpha \approx 1$ the classifiers are dominated by differences in the first coordinate whilst for $\alpha \approx 0$ the classifiers are dominated by differences in the second coordinate. More precisely, let

$$\mu_1(\xi_i) = \begin{cases} 1 & \text{if } \xi_{2i} \leqslant c_1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mu_2(\xi_i) = \begin{cases} 1 & \text{if } \xi_{1i} \leqslant c_2 \\ 0 & \text{otherwise.} \end{cases}$$

Then define

$$\Delta\mathcal{E}_n = \mathcal{E}_n(\mu_1) - \mathcal{E}_n(\mu_2).$$
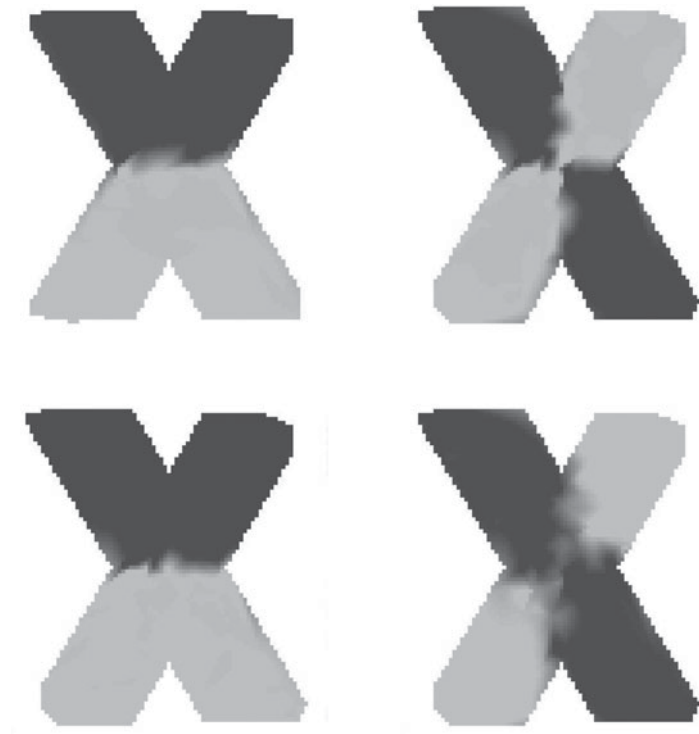
The results are given in figure 3.

Figure 2. The top row shows the minimizers of $\mathcal{E}_n$ and the bottom row shows the minimizers of $GTV_n$ for the graph given in figure 1 conditioned on the node closest to each corner taking either 0 or 1. The left column is conditioned to have 0 in the bottom corners and 1 in the top corners. The right column has 0 in the bottom left and top right corners and 1 in the top left and bottom right corners. There is very little difference between the outputs on the left but on the right the $GTV_n$ term fails to pick out the singularity at the centre.

## 1.3. The limiting model

Rather surprisingly, the problem of soft classifications for finite data sets and hard classification in the limit has received relatively little attention in the literature. However, it is well known that for finite data, one can recover the $k$-means algorithm (hard classification) from the expectation-maximization algorithm (soft classification) in the zero-variance limit for the Gaussian mixture model and the Dirichlet process mixture model [**31**, **35**].

The results of this paper concern the asymptotics of the minimum and minimizers of $\mathcal{E}_n$, where $\epsilon_n \to 0$ as $n \to \infty$. The advantages of scaling $\epsilon_n$ to zero are two-fold. The first is that the matrix $W = (W_{ij})_{ij}$ is sparse and, therefore, we expect the minimization to be numerically less expensive than solving the minimization with a non-sparse matrix (since the sparse minimization has $O(n)$ terms and the non-sparse minimization $O(n^2)$ terms). The second is to improve the resolution of the boundary. One can think of soft classification as estimating the probability that a data point belongs to a certain class and the hard classification problem as
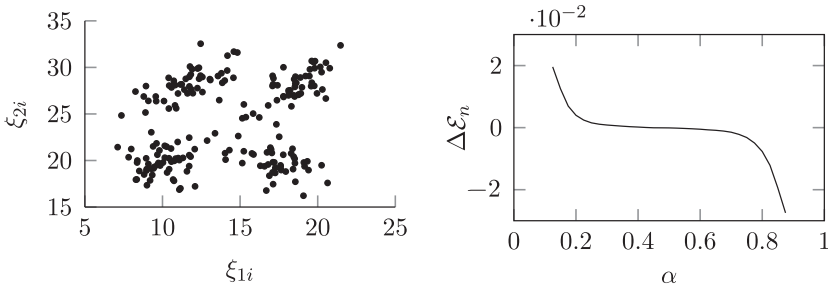
Figure 3. The figure on the left shows a realization of the data. The interactions are parameterized by a potential $\alpha$ which favours horizontal partitions for $\alpha \approx 0$ and vertical partitions for $\alpha \approx 1$ as shown by the figure on the right.
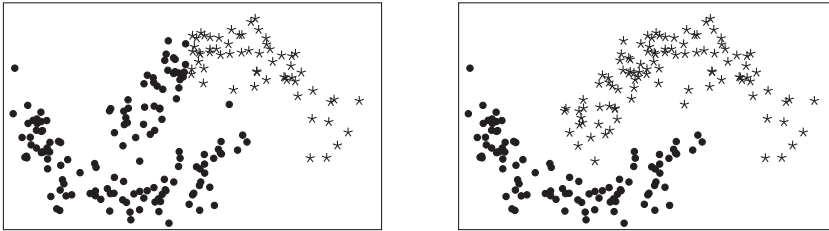


Figure 4. Both figures were classified using the Ginzburg–Landau functional. The one on the left used a larger $\epsilon$ than the one on the right. The figure shows that the smaller value of $\epsilon$ gives a much better resolution in the boundary.

estimating the boundaries where one class is more likely than all others. By scaling $\epsilon_n \to 0$ it will be shown that the limiting minimization problem is a hard classification. For example, figure 4 shows (for a fixed number of data points) improved resolution in the boundary between classes as $\epsilon \to 0$. See also [**24**].

Assume $X \subset \mathbb{R}^d$ and define $\mathcal{E}_\infty : L^1(X) \to [0, \infty]$ by

$$\mathcal{E}_\infty(\mu) = \begin{cases} \int_{\partial\{\mu=1\}} \sigma(n(x))\rho^2(x)\, \mathrm{d}\mathcal{H}^{d-1}(x) & \text{if } \mu \in L^1(X; \{0,1\}) \\ \infty & \text{otherwise} \end{cases} \tag{4}$$

where $n(x)$ is the outward unit normal for the set $\partial\{\mu = 1\}$, $\mathcal{H}^{d-1}$ is the $d-$ one$-$dimensional Hausdorff measure and

$$\sigma(\nu) = \int_{\mathbb{R}^d} \varphi(\pi(x))|x \cdot \nu|\, \mathrm{d}x.$$

It will be shown, in the sense of $\Gamma$-convergence, that $\mathcal{E}_\infty$ is the limiting problem and any sequence such that $\mathcal{E}_n(\mu_n)$ is bounded is precompact in an appropriate topology. In particular, this allows one to apply the results of this paper to infer the consistency of the constrained minimization problem (see § 2.2).

We now briefly discuss the convergence of $\mu_n \to \mu$. Since each $\mu_n$ is defined on a different space (the domain of each $\mu_n$ is $\Psi_n$) it is not straightforward what is meant by the convergence of $\mu_n \to \mu_\infty$. By defining a map $T_n : X \to \Psi_n$ one can compare

$\mu_n$ with $\mu_\infty$ by defining the piecewise constant approximation of $\mu_n$. Formally, we can say $\mu_n \to \mu_\infty$ in $TL^1$ if $\mu_n \circ T_n \to \mu_\infty$ in $L^1$, this is discussed further in § 3.4.

We also include preliminary results towards characterizing the rate of convergence by considering a simple example when $\mu = \mathbb{I}_E$ for a polyhedral set $E \subset X$ and looking at the convergence in the mean square:

$$\mathbb{E} \left| \mathcal{E}_n(\mu) - \mathcal{E}_\infty(\mu) \right|^2 = \mathbb{E} \left| GTV_n(\mu) - TV(\mu; \rho, \eta) \right|^2.$$

We give an expansion of the above in terms of $\epsilon_n$ and $n$. A further overview of these results is given in § 2.3 and the proofs in § 6.

The outline of the paper is as follows. In § 2 the main result, theorem 2.3 (the convergence of the unconstrained minimization problem), is given. We also include an overview of the preliminary rate of convergence results to be found in § 6. Section 3 contains the background material and in particular: notation, a brief overview on $\Gamma$-convergence, background on total variation distances and the key details required from transportation theory. In § 4, the proof of the first part of theorem 2.3 (the compactness result) is given. And in § 5, the proof is completed with the $\Gamma$-convergence result. Finally, in § 6, we make the preliminary calculation regarding the rate of convergence of '$\mathcal{E}_n \to \mathcal{E}_\infty$'.

## 2. Statement of main result and assumptions

The assumptions on $V, \varphi$ and $\pi$ are given in the following definition.

DEFINITION 2.1. *We say that the functions $(\rho, \epsilon_n, V, \varphi, \pi)$ where $\rho : \mathbb{R}^n \to [0, \infty)$, $V : \mathbb{R} \to [0, \infty)$, $\varphi : \mathbb{R} \to [0, \infty)$ and $\pi : \mathbb{R}^d \to \mathbb{R}$ are $\mathcal{E}_n$-admissible if the following conditions hold.*

(1) *The support of $\rho$ is $X$ where $X \subset \mathbb{R}^d$ is open, bounded, connected, Lipschitz boundary and $1 \leqslant d < \infty$.*

(2) *On $X$ we have that $\rho$ is a continuous probability density ($\int_X \rho(x)\, \mathrm{d}x = 1$) and bounded above and below by positive constants, that is,*

$$0 < \inf_{x \in X} \rho(x) \leqslant \sup_{x \in X} \rho(x) < \infty.$$

(3) *$\epsilon_n^{-1} = o(f_d(n))$ where*

$$f_d(n) = \begin{cases} \sqrt{\dfrac{n}{\log \log n}} & \text{if } d = 1 \\[2ex] \dfrac{\sqrt{n}}{(\log n)^{3/2}} & \text{if } d = 2 \\[2ex] \sqrt[d]{\dfrac{n}{\log n}} & \text{if } d \geqslant 3. \end{cases}$$

(4) *The support of $\eta = \varphi \circ \pi$ is compact.*

(5) *$\eta(0) > 0$.*

(6) *For all $\delta > 0$ there exists $c_\delta, \alpha_\delta > 0$ such that if $|x - y| < \delta$ then $\eta(y) \geqslant c_\delta \eta(\alpha_\delta x)$ and furthermore $c_\delta, \alpha_\delta \to 1$ as $\delta \to 0$.*

(7) *$V(y) = 0$ if and only if $y \in \{0, 1\}$.*

(8) *$V$ is continuous.*

(9) *There exists $r > 0$ and $\tau > 0$ such that if $|t| \geqslant r$ then $V(t) \geqslant \tau|t|$.*

The lower bound on $\epsilon_n$ implies the graph with vertices $\Psi_n$ and edges weighted by $W_{ij}$ is (with probability one) eventually connected [**38**, theorem 13.2].

Note that a consequence of 4 and 6 is that $\sigma(\nu) < \infty$ however 4 also excludes any $\pi$ that is linear. For example, if $\pi(x) = w \cdot x$ then there exists $w'$ orthogonal to $w$ such that $w \cdot w' = 0$ (for $d \geqslant 2$). Then $\pi(\alpha w') = 0$ for all $\alpha$ and, in particular, $\eta(\alpha w') = \eta(0) > 0$. Therefore, the support of $\eta$ is not compact. We discuss the linear case more in the following subsection.

Condition 6 gives the required scaling in $\varphi \circ \pi$. If $T_n : X \to \Psi_n$ is a map such that $T_{n\#}P = P_n$ then the continuous approximation of the weights $W_{ij}$ reads as

$$W_{ij} = \frac{1}{\epsilon_n^d} \varphi \circ \pi \left( \frac{\xi_i - \xi_j}{\epsilon_n} \right) \approx \frac{1}{\epsilon_n^d} \int_{T_n^{-1}(\xi_i)} \int_{T_n^{-1}(\xi_j)} \varphi \circ \pi \left( \frac{x - y}{\epsilon_n} \right) \, dx \, dy.$$

Condition 6 is sufficient to formalize this reasoning. We give two different classes of functions in proposition 2.2 below that satisfy the assumptions. The first is a subset of isotropic functions that we extend in corollary 2.4, the second is a set of indicator functions.

PROPOSITION 2.2. *Assume that either*

(1) *$\varphi : [0, \infty) \to [0, \infty)$ is decreasing, $\varphi(0) > 0$, is Lipschitz with compact support and $\pi : \mathbb{R}^d \to [0, \infty)$ is defined by $\pi(x) = |x|$, or*

(2) *$\varphi(0) > 0$, $\varphi(1) = 0$, $\pi : \mathbb{R}^d \to \{0, 1\}$ is given by $\pi(x) = \mathbb{I}_{E^c}(x)$ for an open, bounded and convex set $E \subset \mathbb{R}^d$ with $0 \in E$.*

*Then $(\varphi, \pi)$ are $\mathcal{E}_n$-admissible.*

We prove the above proposition in Appendix A. We now state the main result. The key idea is that optimal transportation theory provides a natural extension of $\mu_n : \Psi_n \to \mathbb{R}$ to $\tilde{\mu}_n : X \to \mathbb{R}$ for which we can use to define the convergence $\mu_n \to \mu$. This is formalized as convergence in $TL^1$, see § 3.4. Establishing $\Gamma$-convergence of functionals, see definition 3.1, and the compactness of minimizers leads to the convergence of minimizers as in theorem 3.2. We use the notation $L^1(\Psi_n)$ as convenient notation for functions defined on $\Psi_n$. We define the space $B(X; \rho, \eta)$, the space of functions of bounded variation with respect to a measure $\rho$ and an interaction potential $\eta$, in definition 3.3.

THEOREM 2.3. *Let $\mathcal{E}_n : L^1(\Psi_n) \to [0, \infty)$ and $\mathcal{E}_\infty : L^1(X) \to [0, \infty]$ be defined by (2) and (4), respectively. If $(\rho, \epsilon_n, V, \varphi, \pi)$ are $\mathcal{E}_n$-admissible (in the sense of definition 2.1) and $\xi_i \overset{\text{iid}}{\sim} \rho$ then, with probability one, the following hold*

(1) *Compactness: Let $\mu_n$ be a sequence of functions on $\Psi_n$ such that $\sup_{n \in \mathbb{N}} \mathcal{E}_n(\mu_n) < \infty$ then $\mu_n$ is relatively compact in $TL^1$ and each cluster point is in $BV(X; \rho, \varphi \circ \pi) \cap L^1(X; \{0, 1\})$.*

(2) *$\Gamma$-limit: We have*

$$\Gamma\text{-}\lim_{n \to \infty} \mathcal{E}_n = \mathcal{E}_\infty.$$

The compactness result is proved in § 4 and the $\Gamma$-convergence result in § 5.

Proposition 2.2 allows one to apply theorem 2.3 to isotropic weights $W_{ij}$ when $\varphi$ is decreasing, $\varphi(0) > 0$ and Lipschitz. We now show that the Lipschitz assumption can be removed.

COROLLARY 2.4. *Let $\mathcal{E}_n : L^1(\Psi_n) \to [0, \infty)$ and $\mathcal{E}_\infty : L^1(X) \to [0, \infty]$ be defined by (2) and (4), respectively. If*

- *$\rho$ satisfies conditions 1–3 in definition 2.1,*

- *$\xi_i \overset{\text{iid}}{\sim} \rho$,*

- *$\varphi : [0, \infty) \to [0, \infty)$ is decreasing, compactly supported, $\varphi(0) > 0$ and continuous at 0,*

- *$\pi(x) = |x|$, and*

- *$V$ satisfies conditions 11–13 in definition 2.1*

*then, with probability one, the conclusions of theorem 2.3 hold.*

We prove the corollary in Appendix B

## 2.1. Convergence of the Ginzburg–Landau functional for linear feature projections

After definition 2.1, we discussed how the compact support assumption on $\eta = \varphi \circ \pi$ did not allow for linear $\pi$. One case that is of interest, that is not covered by theorem 2.3 or corollary 2.4 is $\pi(x) = x_i$ that corresponds to weighting the graph based on differences in one direction only.

This case is of particular interest for high-dimensional data. If the data come from a high, potentially infinite, dimensional space, then it becomes necessary to identify a finite number of principal dimensions upon which to define the edge weights. Isotropic weights are unrealistic in high dimensions and infeasible in infinite dimensions due to the lack of integrability of $\eta$. This motivates our study of linear $\pi$, which can include $\pi(x) = \sum_{i \in \mathcal{K}} w_i x_i$ for a finite set $\mathcal{K}$ and weights $w_i$. Although, we do not consider infinite-dimensional data spaces here, we believe the results of this section can be extended from the finite-dimensional setting, albeit with a modified limit $\hat{\mathcal{E}}_\infty$ to the one we define in (6).

The underlying problem, and why we should not expect a linear choice of $\pi$ to imply that $\mathcal{E}_\infty$ is well defined, is that it becomes harder to ensure that $\sigma(\nu) < \infty$. In particular, we expect $\mathcal{E}_\infty \equiv +\infty$. In many applications, we anticipate that only a

small number of dimensions are relevant and hence we, in this section, consider the classifier as a functional on a projected data space. We define $\hat{\Psi}_n = \{\hat{\xi}_i\}_{i=1}^n \subset \mathbb{R}$ by $\hat{\xi}_i = \pi(\xi_i)$ and $\hat{\mu}_n : \hat{\Psi}_n \to [0, \infty)$. We now consider the energy

$$\hat{\mathcal{E}}_n(\hat{\mu}_n) = \frac{1}{\epsilon_n} \frac{1}{n} \sum_{i=1}^n V(\hat{\mu}_n(\hat{\xi}_i)) + \frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{i,j} \hat{W}_{ij} |\hat{\mu}_n(\hat{\xi}_i) - \hat{\mu}_n(\hat{\xi}_j)| \tag{5}$$

where

$$\hat{W}_{ij} = \frac{1}{\epsilon_n} \varphi \circ \pi \left( \frac{\xi_i - \xi_j}{\epsilon_n} \right) = \frac{1}{\epsilon_n} \varphi \left( \frac{\hat{\xi}_i - \hat{\xi}_j}{\epsilon_n} \right).$$

We point out that the scaling is with respect to $\epsilon_n$ rather than $\epsilon_n^d$ as in (1). In this case, the $\Gamma$-limit $\hat{\mathcal{E}}_\infty : L^1(\mathbb{R}) \to [0, \infty]$ is given by

$$\hat{\mathcal{E}}_\infty(\hat{\mu}) = \hat{\sigma} \sum_{x \in \partial\{\hat{\mu}=1\}} \left( \int_{\pi^{-1}(x)} \rho(y) \, \mathrm{d}y \right)^2 \tag{6}$$

where

$$\hat{\sigma} = \int_{\mathbb{R}} \varphi(x) |x| \, \mathrm{d}x.$$

The analogous assumptions are given in the definition below.

DEFINITION 2.5. *We say that the functions* $(\rho, \epsilon_n, V, \varphi, \pi)$ *where* $\rho : \mathbb{R}^d \to [0, \infty)$, $V : \mathbb{R} \to [0, \infty)$, $\varphi : \mathbb{R} \to [0, \infty)$ *and* $\pi : \mathbb{R}^d \to \mathbb{R}$ *are* $\hat{\mathcal{E}}_n$-*admissible if the following hold.*

(1) *The support of* $\rho$ *is* $X$ *where* $X \subset \mathbb{R}^d$ *and the support of* $\pi_{\#}\rho$ *is open, bounded and connected.*

(2) *On* $X$ *we have that* $\rho$ *is a continuous probability density, bounded above and* $\inf_{x \in X} \pi_{\#}\rho(x) > 0$.

(3) $\epsilon_n^{-1} = o(f_1(n))$ *where* $f_1$ *is given in definition* 2.1.

(4) $\pi$ *is linear.*

(5) *The support of* $\varphi$ *is compact.*

(6) $\varphi(0) > 0$.

(7) *For all* $\delta > 0$ *there exists* $c_\delta, \alpha_\delta > 0$ *such that if* $|x - y| < \delta$ *then* $\varphi(y) \geqslant c_\delta \varphi(\alpha_\delta x)$ *and furthermore* $c_\delta, \alpha_\delta \to 1$ *as* $\delta \to 0$.

(8) $V(y) = 0$ *if and only if* $y \in \{0, 1\}$.

(9) $V$ *is continuous.*

(10) *There exists* $r > 0$ *and* $\tau > 0$ *such that if* $|t| \geqslant r$ *then* $V(t) \geqslant \tau |t|$.

The convergence theorem is given below.

THEOREM 2.6. *Let* $\hat{\Psi}_n = \{\hat{\xi}_i\}_{i=1}^n$ *where* $\hat{\xi}_i = \pi(\xi_i)$. *Define* $\hat{\mathcal{E}}_n : L^1(\hat{\Psi}_n) \to [0, \infty)$ *and* $\hat{\mathcal{E}}_\infty : L^1(X) \to [0, \infty]$ *by* (5) *and* (6), *respectively. If* $(\rho, \epsilon_n, V, \varphi)$ *are* $\hat{\mathcal{E}}_n$-*admissible (in the sense of definition* 2.5) *and* $\xi_i \overset{\text{iid}}{\sim} \rho$ *then, with probability one, the following hold*

(1) *Compactness: Let* $\hat{\mu}_n$ *be a sequence of functions on* $\hat{\Psi}_n$ *such that* $\sup_{n \in \mathbb{N}} \hat{\mathcal{E}}_n(\hat{\mu}_n) < \infty$ *then* $\hat{\mu}_n$ *is relatively compact in* $TL^1$ *and each cluster point is in* $BV(\mathbb{R}; \pi_{\#}\rho, \eta) \cap L^1(\mathbb{R}; \{0, 1\})$.

(2) $\Gamma$-*limit: we have*

$$\Gamma\text{-}\lim_{n \to \infty} \hat{\mathcal{E}}_n = \hat{\mathcal{E}}_\infty.$$

The proof is an application of theorem 2.3 to the one-dimensional data set $\hat{\Psi}_n$.

## 2.2. Comments on the main result

The classical Ginzburg–Landau functional:

$$F_\epsilon(\mu) = \epsilon \int_X |\nabla \mu(x)|^2 \, \mathrm{d}x + \frac{1}{\epsilon} \int_X V(\mu(x)) \, \mathrm{d}x$$

has been well studied and its convergence to a total variation functional

$$F_\infty(\mu) = \sigma_V \int_X |\nabla \mu(x)| \, \mathrm{d}x$$

known for some time [37] and similar results for the anisotropic version [1]. More recent results have studied this functional on a (deterministic) regular graph. In [42] the authors show the $\Gamma$-convergence and compactness of two variants of the Ginzburg–Landau functional where $\{\xi_i\}_{i=1}^n \subset \mathbb{R}^2$ form a 4-regular graph. Let us exploit the structure of the graph by writing data as $\{\xi_{i,j}\}_{i,j=1}^n$ where $\xi_{i,j}, \xi_{i,j+1}$ are neighbours, as are $\xi_{i,j}$ and $\xi_{j+1,i}$. The two variants of the Ginzburg–Landau functional considered in [42] are

$$h_{n,\epsilon}(\mu) = \frac{1}{\epsilon} \sum_{i,j=1}^n V(\mu(\xi_{i,j}))$$

$$+ \frac{1}{n} \sum_{i,j=1}^n \left( |\mu(\xi_{i+1,j}) - \mu(\xi_{i,j})|^2 + |\mu(\xi_{i,j+1}) - \mu(\xi_{i,j})|^2 \right)$$

$$k_{n,\epsilon}(\mu) = \frac{1}{\epsilon n^2} \sum_{i,j=1}^n V(\mu(\xi_{i,j}))$$

$$+ \epsilon \sum_{i,j=1}^n \left( |\mu(\xi_{i+1,j}) - \mu(\xi_{i,j})|^2 + |\mu(\xi_{i,j+1}) - \mu(\xi_{i,j})|^2 \right).$$

The first functional $h_{n,\epsilon}$ $\Gamma$-converges as $\epsilon \to 0$ (for a fixed $n$) to a total variation function $h_{n,0}$ in a discrete setting defined by

$$h_{n,0}(\mu) = \begin{cases} \dfrac{1}{n} \displaystyle\sum_{i,j=1}^{n} \Big( |\mu(\xi_{i+1,j}) - \mu(\xi_{i,j})|^2 \\ \qquad\qquad + |\mu(\xi_{i,j+1}) - \mu(\xi_{i,j})|^2 \Big) & \text{if } \mu \in L^1(\Psi_n; \{0,1\}) \\ \infty & \text{otherwise.} \end{cases}$$

As $\epsilon \to 0$ and $n \to \infty$ sequentially or for $\epsilon = n^{-\alpha}$ for $\alpha$ within some range then $h_{n,\epsilon}$ $\Gamma$-converges to an anisotropic total variation in a continuous setting

$$\Gamma\text{-}\lim_{\substack{n \to \infty \\ \epsilon \to 0}} h_{n,\epsilon} = \int_{\mathbb{T}^2} \left| \frac{\partial \mu}{\partial x} \right| + \left| \frac{\partial \mu}{\partial y} \right|$$

and $k_{n,\epsilon}$ $\Gamma$-converges to an isotropic total variation

$$\Gamma\text{-}\lim_{\substack{n \to \infty \\ \epsilon \to 0}} k_{n,\epsilon} = \int_{\mathbb{T}^2} |\nabla \mu|$$

upto renormalization. Also discussed in [**42**] is the application to the constrained minimization problem

In the remainder of the paper, we will work on point clouds, that are random graphs, rather than deterministic regular graphs.

*Convergence of the Graph Total Variation.* The Ginzburg–Landau functional consists of two terms, the first is a regularization on the derivative and the second is a penalization on states outside of $\{0,1\}$. In this paper, we use the graph total variation on graphs (3) as the regularization on the derivative. In more generality, one can define the $p$-Laplacian on graphs [**49**] by

$$J_p(\mu) = \frac{1}{\epsilon_n^p} \frac{1}{n^2} \sum_{i,j} W_{ij} |\mu(\xi_i) - \mu(\xi_j)|^p.$$

The graph total variation corresponds to $p = 1$ and the results of [**42**] described above correspond to $p = 2$.

An interesting and important question in its own right is the convergence of the $p$-Laplacian [**3**, **11**, **13**, **19**, **48**, **50**]. For isotropic weights, that is, $\pi(x) = |x|$, the following result was established for $p = 1$ in [**25**].

(1) Compactness: Let $\mu_n$ be any sequence of functions on $\Psi_n = \{\xi_i\}$, where $\xi_i \overset{\text{iid}}{\sim} P$, such that $GTV_n(\mu_n)$ is bounded and $\mu_n$ is bounded in $TL^1$ then $\mu_n$ is almost surely relatively compact in $TL^1$ and each cluster point is in $BV(X; \rho, \eta)$.

(2) $\Gamma$-limit: we have, with probability one,

$$\Gamma\text{-}\lim_{n \to \infty} GTV_n = TV(\cdot; \rho, \eta).$$

A consequence of our proofs is that, this result is also true for any $X, \rho, \epsilon_n, \varphi$ and $\pi$ satisfying assumptions 1–6 in definition 2.1. In particular, the $\Gamma$-convergence holds analogously to theorem 5.1 and the compactness can be reduced, as in the proof of Proposition 4.1, to the isotropic case where the results of [25] apply. This generalizes the result of [25] to anisotropic weights.

*Convergence of minimizers.* The results of the theorem 2.3 can be understood as implying the convergence of minimizers in the following sense. For a sequence of closed sets $\Theta_n \subseteq L^1(\Psi_n)$ and $\Theta \subseteq L^1(X)$ which we assume with respect to the $\Gamma$-convergence, that is, the following hold: (1) if $\zeta \in \Theta$ then $\zeta_n := \zeta\big|_{\Psi_n} \in \Theta_n$, (2) there exists $\zeta \in \Theta$ such that $\mathcal{E}_\infty(\zeta) < \infty$ and (3) any sequence $\zeta_n$ with $\zeta_n \to \zeta$ implies $\zeta \in \Theta$. With probability one the following statements hold.

(1) Convergence of the minimum: $\lim_{n\to\infty} \inf_{\Theta_n} \mathcal{E}_n = \min_\Theta \mathcal{E}_\infty$.

(2) Convergence of minimizers: if $\mu_n \in L^1(\Psi_n)$ are a sequence satisfying

$$\mathcal{E}_n(\mu_n) \leqslant \min_\mu \mathcal{E}_n(\mu) + \delta_n$$

for a sequence $\delta_n \to^+ 0$ (we call $\mu_n$ a minimizing sequence) then this sequence is precompact in $TL^1$ and furthermore, any cluster point minimizes $\mathcal{E}_\infty$.

The proof is a simple consequence of theorems 2.3 and 3.2.

Alternatively, one could let $g_n : L^1(\Psi_n) \to [0, \infty)$ be a sequence that continuously converges to $g : L^1(X) \to [0, \infty)$, that is, $g_n(\zeta_n) \to g(\zeta)$ whenever $\zeta_n \to \zeta$ in $L^1$, and then since the $\Gamma$-convergence is stable under continuous perturbations the results of this paper imply (with probability one) that

(1) $\lim_{n\to\infty} \inf_{L^1(\Psi_n)} (\mathcal{E}_n + g_n) = \min_{L^1(X)} (\mathcal{E}_\infty + g)$, and

(2) if $\mu_n \in L^1(\Psi_n)$ are a minimizing sequence for $\mathcal{E}_n + g_n$ then this sequence is precompact in $TL^1$ and furthermore, any cluster point minimizes $\mathcal{E}_\infty + g$.

For example, one could use this in order to fit data, for example,

$$g_n(\mu; \zeta) = \lambda \int_X |\mu(T_n(x)) - \zeta(x)| \, dx$$

where $\zeta$ is a known function (data) and $T_n$ is a sequence of stagnating transport maps (a sequence such that $\|\mathrm{Id} - T_n\|_{L^p} \to 0$, see § 3.4). In this case, $g(\mu) = \lambda \int_X |\mu(x) - \zeta(x)| \, dx$.

*Choice of scaling.* The natural choice of scaling in $\mathcal{E}_n$ between the two terms is not a-priori obvious. One could write

$$\mathcal{E}_n(\mu) = \frac{1}{\gamma_n} \sum_{i=1}^n V(\mu(\xi_i)) + \frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{i,j} W_{ij} |\mu(\xi_i) - \mu(\xi_j)|.$$

The proof of Theorem 4.1 requires $\gamma_n/\epsilon_n = O(1)$. One can show that theorem 2.3 holds for $\gamma_n = O(\epsilon_n)$. For simplicity, it is assumed that $\gamma_n = \epsilon_n$.

*Generalization to $L^p$ spaces for $p > 1$.* The literature on the related problem in $L^p$ spaces has been well developed. Define

$$\mathcal{E}_n(\mu) = \frac{1}{\epsilon_n} \sum_{i=1}^{n} V(\mu(\xi_i)) + \frac{1}{\epsilon_n^p} \frac{1}{n^2} \sum_{i,j} W_{ij} \left| \mu(\xi_i) - \mu(\xi_j) \right|^p.$$

Then one can show the results of theorem 2.3 still hold with a modification in the integrand of the limiting energy $\mathcal{E}_\infty$; in particular, the limit is still a total variation type semi-norm on binary valued functions. In particular, the convergence of minimizing sequences is still in $TL^1$.

*Size of the phase transition.* Although we do not formally state the result, it is well known that for the Ginzburg–Landau functional the phase transition is of order $\epsilon_n$, see for example [8, theorem 6.4]. By the proof of Lemma 5.4 the same is true in the setting described in this paper. That is, if $\mu_n$ is a recovery sequence (see definition 3.1) for $\mathcal{E}_n$ then $1/n \#\{\xi \in \Psi_n : \mu_n(\xi) \in (c, 1-c)\} = O(\epsilon_n)$.

## 2.3. Preliminary results on the rate of convergence

We include some preliminary results concerning the rate of convergence for $\inf \mathcal{E}_n \to \min \mathcal{E}_\infty$. The problem is simplified by looking at the convergence $GTV_n(\mu) \to TV(\mu; \rho, \eta)$ for $\mu = \mathbb{I}_E$ where $E$ is a polyhedral set. To characterize the rate of convergence, we look for convergence in mean square. It is shown in theorem 6.2 that

$$\mathbb{E} \left| \mathcal{E}_n(\mu) - \mathcal{E}_\infty(\mu) \right|^2 = \mathbb{E} \left| GTV_n(\mu) - TV(\mu; \rho, \eta) \right|^2$$

$$= O(\epsilon_n) + \frac{\kappa_1}{n\epsilon_n} + \frac{\kappa_2}{n^2 \epsilon_n^{d+1}} + O\left(\epsilon_n^2\right)$$

as $n \to \infty$ for a constants $\kappa_i$ given in § 6. Even though (by lemma 5.4) one has $GTV_n(\mu) \to TV(\mu; \rho, \eta)$ almost surely convergence in expectation does not immediately follow, this is shown in theorem 6.1. The leading $O(\epsilon_n)$ term above corresponds to approximating $\mu$ along edges of $E$ where an edge is the intersection of two faces of $E$ (see § 6 for a precise explanation of the notation we use to describe polyhedral sets). The error in the edges causes a bias in the estimate. For example, if one considers the function $\mu = \mathbb{I}_{H \cap X}$ where $H = \{x : w \cdot x > 0\}$ for some $w$ is any half space then $\mu$ is a polyhedral function with no edges in $X$ and one can show $\mathbb{E} GTV_n(\mu) = TV(\mu; \rho, \eta)$. It follows from our proofs that $\mathbb{E} \left| GTV_n(\mu) - TV(\mu; \rho, \eta) \right|^2 = \kappa_1/n\epsilon_n + ((\kappa_2)/(n^2 \epsilon_n^{d+1})) + O(1/n)$.

## 3. Preliminary material

### 3.1. Notation

The space of functions from $Z$ onto $Y$ that are $L^p$-integrable are denoted by $L^p(Z; Y)$ (for $1 \leqslant p \leqslant \infty$). Usually, either $Y = \{0, 1\}$ or $Y = \mathbb{R}$. If $Y = \mathbb{R}$ then we write $L^p(Z)$ instead of $L^p(Z; \mathbb{R})$. When we use the $L^p$ norm with respect to a measure $P$ the $Y$ dependence is suppressed and we write $L^p(X; P)$. It will be obvious from the context what is meant.

The Euclidean norm is given by $|\cdot|$ and with a small abuse of notation the dimension is inferred from the argument. The ball centred at $x$ and with radius $r$ in $\mathbb{R}^d$ is given as

$$B(x, r) = \left\{ y \in \mathbb{R}^d : |x - y| < r \right\}.$$

When the ball is centred at the origin, we write $B(0, r)$.

### 3.2. Γ-Convergence

Γ-convergence was introduced in the 1970s by De Giorgi as a tool for studying sequences of variational problems. A key feature is the natural emergence of singular objects such as characteristic functions which we will interpret as classifiers. We will adopt the viewpoint that Γ-convergence can used as a data analysis tool as it provides direct links between statistical modelling assumptions and the corresponding minimizers, cf. for example, § 1.2. Very accessible introductions to Γ convergence can be found in [**8**, **17**].

We have the following definition of Γ-convergence.

DEFINITION 3.1 Γ-convergence. *Let $(A, d_A)$ be a metric space. A sequence $f_n : A \to \mathbb{R} \cup \{\pm\infty\}$ is said to Γ-converge on the domain $A$ to $f_\infty : A \to \mathbb{R} \cup \{\pm\infty\}$ with respect to $d_A$, and write $f_\infty = \Gamma\text{-}\lim_n f_n$, if for all $\zeta \in A$:*

(i) *(lim inf inequality) for every sequence $(\zeta_n)$ converging to $\zeta$*

$$f_\infty(\zeta) \leqslant \liminf_{n\to\infty} f_n(\zeta_n);$$

(ii) *(recovery sequence) there exists a sequence $(\zeta_n)$ converging to $\zeta$ such that*

$$f_\infty(\zeta) \geqslant \limsup_{n\to\infty} f_n(\zeta_n).$$

When it exists the Γ-limit is always lower semi-continuous, and hence there exists minimizers over compact sets. The following result justifies the use of Γ-convergence as a variational type of convergence.

THEOREM 3.2 Convergence of Minimizers. *Let $(A, d_A)$ be a metric space and $f_n : A \to [0, \infty]$ be a sequence of functionals. Let $\mu_n$ be a minimizing sequence for $f_n$. If $\mu_n$ are precompact and $f_\infty = \Gamma\text{-}\lim_n f_n$ where $f_\infty : A \to [0, \infty]$ is not identically $+\infty$ then*

$$\min_A f_\infty = \lim_{n\to\infty} \inf_A f_n.$$

*Furthermore, any cluster point of $\mu_n$ minimizes $f_\infty$.*

A simple consequence of the above is if one can show that the Γ-limit has a unique minimizer then any minimizing sequence converges (without the recourse to subsequences).

### 3.3. Total variation distance

The energy $\mathcal{E}_\infty$ can also be written as a BV norm:

$$\mathcal{E}_\infty(\mu) = \begin{cases} TV(\mu; \rho, \eta) & \text{if } \mu \in L^1(X; \{0, 1\}) \\ \infty & \text{otherwise} \end{cases} \tag{7}$$

where

$$TV(\mu; \rho, \eta) = \sup\left\{ \int_X \mu(x) \operatorname{div}(\phi(x)) \, dx \ : \ \phi \in C_c^\infty(X; \mathbb{R}^d), \right.$$

$$\left. \sup_{x \in X} \sigma^* \left( -\rho^{-2}(x)\phi(x) \right) < \infty \right\}, \tag{8}$$

$$\sigma^*(\phi) = \sup\left\{ \nu \cdot \phi - \sigma(\nu) \ : \ \nu \in \mathbb{R}^d \right\} \in \{0, \infty\}, \tag{9}$$

$$\sigma(\nu) = \int_{\mathbb{R}^d} \eta(x)|x \cdot \nu| \, dx, \tag{10}$$

$\eta = \varphi \circ \pi$ and $\rho$ is a probability density on $X$. The surface energy density $\sigma$ is 1-homogeneous and the Legendre transform $\sigma^*$ is a characteristic function which assumes the values 0 or $\infty$. The space of functions with bounded variation is defined below.

DEFINITION 3.3. *For a domain $X \subset \mathbb{R}^d$ the weighted total variation $TV(\cdot; \rho, \eta)$ of function $\mu \in L^1(X)$ with respect to a density $\rho$ and potential $\eta$ is defined by (8–10). The space of functions with finite weighted total variation is denoted by $BV(X; \rho, \eta)$. The standard total variation distance on $X$ is defined by*

$$\widehat{TV}(\mu) = \sup\left\{ \int_X \mu(x) \, div(\phi) \, dx \ : \ \phi \in C_c^\infty(X), \|\phi\|_{L^\infty(X)} \leqslant 1 \right\}.$$

*The standard bounded variation space $\widehat{BV}(X)$ is the set of functions such that $\widehat{TV}(\mu) < \infty$.*

The equivalence of definitions (4) and (7) can be seen from the simplification of $TV(\cdot; \rho, \eta)$ when $\mu \in C^1$:

$$TV(\mu; \rho, \eta) = \int_X \sigma(\nabla\mu(x))\rho^2(x) \, dx = \int_X \int_{\mathbb{R}^d} \eta(y)|y \cdot \nabla\mu(x)|\rho^2(x) \, dy \, dx.$$

One may also write

$$TV(\mu; \rho, \eta) = \int_{\mathbb{R}^d} \eta(z) TV_z(\mu; \rho) \, dz$$

where $TV_z(\cdot; \rho)$ is defined by

$$TV_z(\mu; \rho) = \sup\left\{ \int_X \mu(x) \operatorname{div}(\phi(x)) \, dx \ : \ \phi \in C_c^\infty(X; \mathbb{R}^d), \right.$$

$$\left. -\nu \cdot \phi(x) \leqslant |z \cdot \nu|\rho^2(x) \, \forall \nu, x \in \mathbb{R}^d \right\}$$

The following proposition is a slight generalization of a well-known result regarding the convergence of difference quotients to the total variation semi-norm. The proof is omitted but it is a trivial adaptation of, for example, [**32**, theorem 13.48].

PROPOSITION 3.4. *Assume $\mu_n \to \mu$ in $L^1$. For a sequence $\epsilon_n \to 0$ and a function $\rho : X \to [0, \infty)$ define $f_n : X \to [0, \infty)$ by*

$$f_n(z) = \frac{1}{\epsilon_n} \int_X |\mu_n(x + \epsilon_n z) - \mu_n(x)| \, \rho^2(x) \, \mathrm{d}x.$$

*Then*

$$\liminf_{n \to \infty} f_n(z) \geqslant TV_z(\mu; \rho).$$

For each $\mu \in BV(X; \rho, \eta)$, the following theorem gives the existence of a measure that one can understand as the weak derivative of $\mu$. See for example [**4**, **20**] for more details.

THEOREM 3.5. *For every $\mu \in BV(X; \rho, \eta)$ there exists a Radon measure $\lambda_{\rho,\eta}$ on $X$ and a $\lambda_{\rho,\eta}$-measurable function $\alpha : X \to \mathbb{R}$ such that $\alpha(x) = 1$ for $\lambda_{\rho,\eta}$-almost every $x \in X$ and*

$$\int_X \mu(x) \mathrm{div} \phi(x) \, \mathrm{d}x = - \int_X \frac{\phi(x) \cdot x}{\rho^2(x)\sigma(x)} \alpha(x) \, \lambda_{\rho,\eta}(\mathrm{d}x)$$

*for all $\phi \in C_c^1(X; \mathbb{R}^d)$. In particular,*

$$\lambda_{\rho,\eta}(X) = TV(\mu; \rho, \eta).$$

*For the standard total variation distance, we write $\hat{\lambda}$ and have the following relationship:*

$$\lambda_{\rho,\eta}(\mathrm{d}x) = \rho^2(x)\sigma(x) \, \hat{\lambda}(\mathrm{d}x).$$

*In particular,*

$$TV(\mu; \rho, \eta) = \int_X \rho^2(x)\sigma(x) \, \hat{\lambda}(\mathrm{d}x). \tag{11}$$

Using (11), we first prove theorem 2.3 (in particular, the $\Gamma$-convergence statement) for Lipschitz $\rho$ then generalize to continuous $\rho$ by taking a monotonic sequence of Lipschitz functions $\rho_k \to \rho$ and applying the monotone convergence theorem.

A useful approximation result we will make use of is for all $\mu \in BV(X; \rho, \eta)$ there exists a sequence $\mu_n \in BV(X; \rho, \eta) \cap C^\infty(\mathbb{R}^d)$ such that $\mu_n \to \mu$ in $L^1(X)$ and $TV(\mu_n; \rho, \eta) \to TV(\mu; \rho, \eta)$ or equivalently $\lambda_n^{(\rho,\eta)}(X) \to \lambda^{(\rho,\eta)}(X)$ (where $\lambda_n^{(\rho,\eta)}$ is the measure given by theorem 3.5 and induced by $\mu_n$), see for example, [**20**, theorem 2, §5.2.2].

### 3.4. Transportation theory

This subsection contains the preliminary material required to compare functions defined on different domains. The key idea in [**25**] was to use transportation maps in order to extend functions to a common domain. This allows one to define an $L^p$-type convergence for functions on different domains.

Let us consider a sequence of functions $\mu_n$ on discrete domains $\Psi_n$ and a function $\nu$ on the continuous domain $X$. In order to compare $\mu_n$ with $\nu$, we first use a piecewise constant approximation $\tilde{\mu}_n$ to extend $\mu_n$ onto the space $X$. For a map $T_n : X \to \Psi_n$, we define $\tilde{\mu}_n = \mu_n \circ T_n$. One can then define a topology using the $L^p$ distance between $\tilde{\mu}_n$ and $\nu$. The challenge is to define $T_n$ optimally in the sense that as little mass as possible is moved. In this section, we have given an overview of the framework proposed in [**25**] that describes the topology we use in the sequel. We start by defining the $p$-OT distance.

DEFINITION 3.6. *If $1 \leqslant p < \infty$ then the $p$-OT distance between $P, Q \in \mathcal{P}(X)$ is defined by*

$$d_p(P,Q) = \min\left\{ \left( \int_{X^2} |x-y|^p \, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p} : \pi \in \Gamma(P,Q) \right\} \qquad (12)$$

*where $\Gamma(P,Q)$ is the set of couplings between $P$ and $Q$, that is, the set of probability measures on $X \times X$ such that the first marginal is $P$ and the second marginal is $Q$.*

*If $p = \infty$ then the $\infty$-OT distance between $P, Q \in \mathcal{P}(X)$ is defined by*

$$d_\infty(P,Q) = \min\left\{ \operatorname*{ess\,sup}_{\pi} \{|x-y| : (x,y) \in X \times X\} : \pi \in \Gamma(P,Q) \right\}. \qquad (13)$$

The minimization problem in (12) and (13) is known as Kantorovich's optimal transportation problem. The minimization is convex and therefore, the minimum is achieved [**16**, **45**]. One can also show that $d_p$ defines a metric on the space of probability measures. Elements $\pi \in \Gamma(P,Q)$ are called transference plans. The distance $d_2$ is also known as the Wasserstein metric and $d_\infty$ the $\infty$-transportation distance. For bounded $X \subset \mathbb{R}^d$ convergence in $d_p$ (for $1 \leqslant p < \infty$) is equivalent to the weak convergence of probability measures [**45**] and therefore, with probability one, $d_p(P_n, P) \to 0$ where $P_n$ is the empirical measure.

When $P$ has density with respect to the Lebesgue measure then the Kantorovich minimization problem is equivalent to the Monge optimal transportation problem [**23**]:

Minimize $\int_X |x - T(x)|^p \, P(\mathrm{d}x)$   overall measurable maps $T$ such that $T_\# P = Q$

where the push forward measure is defined by

$$T_\# P(A) = P(T^{-1}(A))$$

for any $A \in \mathcal{B}(X)$. If $Q = T_\# P$ then we call $T$ a transportation map between $P$ and $Q$.

Let $P_n$ be the empirical measure and $1 \leqslant p < \infty$ then from $d_p(P_n, P) \to 0$ (almost surely) one can immediately infer the existence of a sequence of transport plans such that

$$\|\mathrm{Id} - T_n\|^p_{L^p(X;P)} = \int_X |x - T_n(x)|^p \, P(\mathrm{d}x) \to 0 \tag{14}$$

as $n \to \infty$. We call any sequence of transportation maps $\{T_n\}$ that satisfy (14) *stagnating*.

In the next definition, we use stagnating transport maps to define piecewise constant approximations of functions on $\Psi_n$ in order to define a suitable notion of convergence.

DEFINITION 3.7. *Let* $\mu_n \in L^p(\Psi_n) = L^p(X; P_n)$ *and* $\mu \in L^p(X; P)$ *where* $P$ *is a probability measure and* $P_n = 1/n \sum_{i=1}^n \delta_{\xi_i}$ *is the empirical measure. We say* $\mu_n \to \mu$ *in* $TL^p(X)$ *if*

$$\|\mu_n \circ T_n - \mu\|^p_{L^p(X;P)} = \int_X |\mu_n(T_n(x)) - \mu(x)|^p \, P(\mathrm{d}x) \to 0 \tag{15}$$

*for any sequence of stagnating transportation maps* $T_n : X \to \Psi_n$. *Similarly,* $\mu_n$ *is bounded in* $TL^p$ *if* $\|\mu_n \circ T_n\|_{L^p}$ *is bounded and* $\mu_n$ *is precompact in* $TL^p$ *if* $\mu_n \circ T_n$ *is precompact in* $L^p$.

One can show that if (15) holds for one sequence of stagnating transport maps then it holds for any sequence of stagnating transport maps [**25**, lemma 3.5].

Since it is assumed that $P$ has density $\rho$ which is bounded above and below by positive constants then (14) is equivalent to $\|\mathrm{Id} - T_n\|_{L^p(X)} \to 0$ and $L^p(X; P) = L^p(X)$. This paper focuses on the case where $p = 1$ however, it is straightforward to consider the case $1 \leqslant p < \infty$, see § 2.2.

Now consider an arbitrary $T : X \to X$ and a measurable $\varphi : X \to [0, \infty)$. Recall that

$$\int_X \varphi(x) \, T_\# P(\mathrm{d}x) := \sup \left\{ \int_X s(x) T_\# P(\mathrm{d}x) : 0 \leqslant s \leqslant \varphi \text{ and } s \text{ is simple} \right\}.$$

If $s(x) = \sum_{i=1}^N a_i \delta_{U_i}(x)$ where $a_i = s(x)$ for any $x \in U_i$ then

$$\int_X s(x) T_\# P(\mathrm{d}x) = \sum_{i=1}^N a_i T_\# P(U_i) = \sum_{i=1}^N a_i P(V_i)$$

for $V_i = T^{-1}(U_i)$. Note that $a_i = s(x)$ for any $x \in T(U_i)$. From this it is not hard to see the following change of variables formula:

$$\int_X \varphi(x) \, T_\# P(\mathrm{d}x) = \int_X \varphi(T(x)) \, P(\mathrm{d}x). \tag{16}$$

A particularly useful version of this will be when $T_\# P(\mathrm{d}x) = P_n(\mathrm{d}x)$ where $P_n$ is the empirical measure. In which case (16) implies

$$\frac{1}{n} \sum_{i=1}^n \varphi(\xi_i) = \int_X \varphi(T(x)) \, P(\mathrm{d}x).$$

As an aside, one can generalize the $TL^p$ framework to pairs $(\mu, P)$ where $\mu \in L^p(X; P)$. Let us define

$$d_{TL^p}((P, \mu), (Q, \zeta)) = \inf_{\pi \in \Gamma(P,Q)} \left\{ \left( \int_{X \times Y} |x - y|^p \, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p} \right.$$
$$\left. + \left( \int_{X \times Y} |\mu(x) - \zeta(y)|^p \, \pi(\mathrm{d}x, \mathrm{d}y) \right)^{1/p} \right\}.$$

Let $P$ have density with respect to the Lebesgue measure and take a sequence of measures $P_n$ defined on a common space $X$ (where we do not assume that $P_n$ is the empirical measure). Then $(P_n, f_n) \to (P, f)$ in $TL^p$ is equivalent to weak convergence of measures (due to the first term) and $\|\mu_n \circ T_n - \mu\|_{L^p(X;P)} \to 0$ (due to the second term), see [**25**, proposition 3.6]. Since we are working with the empirical measure then with probability one $P_n$ converges weakly to $P$. Hence the first term plays no role in this paper and so is not included.

Our proofs require a bound on the supremum norm of $T_n - \mathrm{Id}$ given by the following theorem.

THEOREM 3.8. *Let $X \subset \mathbb{R}^d$ with $d \geqslant 1$ be open, connected and bounded with Lipschitz boundary. Let $P$ be a probability measure on $X$ with density (with respect to Lebesgue) $\rho$ which is bounded above and below by positive constants. Let $\xi_1, \xi_2, \ldots$ be a sequence of independent random variables with distribution $P$ and let $P_n$ be the empirical measure. Then there exists a constant $C > 0$ such that almost surely there exists a sequence of transportation maps $\{T_n\}_{n=1}^\infty$ from $P$ to $P_n$ such that*

$$\limsup_{n \to \infty} \frac{\|T_n - \mathrm{Id}\|_{L^\infty(X)}}{\delta_n} \leqslant C$$

*where*

$$\delta_n = \begin{cases} \dfrac{\sqrt{\log \log n}}{\sqrt{n}} & \text{if } d = 1 \\[2mm] \dfrac{(\log n)^{3/4}}{\sqrt{n}} & \text{if } d = 2 \\[2mm] \dfrac{(\log n)^{1/d}}{n^{1/d}} & \text{if } d \geqslant 3. \end{cases}$$

*Proof.* The proof for $d = 2$ and $d \geqslant 3$ can be found in [**26**]. For $d = 1$ the result follows from considering the transportation map defined by $T_n(x) = \xi_i$ for $x \in (x_n^{(i-1)}, x_n^{(i)}]$ where $x_n^{(i)}$ is defined by $P((-\infty, x_n^{(i)}]) = i/n$ for $i = 1, \ldots, n - 1$ and $x_n^{(0)} = -\infty$ and $x_n^{(n)} = \infty$. One then has $\|T_n - \mathrm{Id}\|_{L^\infty(X)} = \|P_n - P\|_{L^\infty(X)}$ which by the law of the iterated logarithm has the stated rate of convergence. $\quad\square$

## 4. The compactness property

In this section, we prove the first part of theorem 2.3 by establishing that sequences bounded in $\mathcal{E}_n$ are precompact in $TL^1$ with cluster points in $L^1(X; \{0, 1\})$. Our

proofs compare $\mathcal{E}_n$ with its continuous analogue $\mathcal{C}_\epsilon : L^1(X) \to [0, \infty]$ defined by

$$\mathcal{C}_\epsilon(\mu) = \frac{1}{\epsilon} \int_X V(\mu(x)) \rho(x) \, dx$$

$$+ \frac{1}{\epsilon^{d+1}} \int_{X^2} \varphi \circ \pi \left( \frac{x-y}{\epsilon} \right) |\mu(x) - \mu(y)| \, \rho(x) \rho(y) \, dx \, dy. \qquad (17)$$

The transport map $T_n$ between the measures $P_n$ and $P$ is used to compare a function $\mu_n : \Psi_n \to \mathbb{R}$ to its continuous version $\tilde{\mu}_n : X \to \mathbb{R}$, i.e. $\tilde{\mu}_n = \mu_n \circ T_n$. One then uses standard results to conclude the compactness of $\tilde{\mu}_n$ in $L^1$ and show that this implies compactness of $\mu_n$ in $TL^1$.

PROPOSITION 4.1. *Under the same conditions as theorem* 2.3. *If* $\mu_n \in L^1(\Psi_n)$ *is a sequence with*

$$\sup_{n \in \mathbb{N}} \mathcal{E}_n(\mu_n) < \infty$$

*then, with probability one, there exists a subsequence* $\mu_{n_m}$ *and* $\mu \in L^1(X; \{0, 1\})$ *such that* $\mu_{n_m} \to \mu$ *in* $TL^1$.

*Proof.* Recall the following preliminary compactness result. If $\{\tilde{\mu}_n\}_{n=1}^\infty$ is a sequence in $L^1(X)$ such that

$$\sup_{n \in \mathbb{N}} \mathcal{C}_{\epsilon_n}(\tilde{\mu}_n) < \infty, \qquad (18)$$

where $\mathcal{C}_{\epsilon_n} : L^1(X) \to [0, \infty]$ is defined by (17), then there exists a subsequence $\tilde{\mu}_{n_m}$ and $\mu \in L^1(X; \{0, 1\})$ such that $\tilde{\mu}_{n_m} \to \mu$ in $L^1$. A proof can be found, for example, in [1].

For clarity, we will denote the dependence of $\eta = \varphi \circ \pi$ on $\mathcal{E}_n$ by $\mathcal{E}_n(\cdot; \eta)$ and let $\eta_\epsilon(x) = 1/\epsilon^d \eta(x/\epsilon)$. Since $\eta$ is continuous at 0 and $\eta(0) > 0$ there exists $b > 0$ and $a > 0$ such that $\eta(x) \geqslant a$ for all $|x| < b$. Define $\tilde{\eta}$ by $\tilde{\eta}(x) = a$ for $|x| < b$ and $\tilde{\eta}(x) = 0$ otherwise. As $\tilde{\eta} \leqslant \eta$ then $\mathcal{E}_n(\mu_n; \eta) \geqslant \mathcal{E}_n(\mu_n; \tilde{\eta})$.

Let $T_n$ be such that $T_{n\#}P = P_n$ and the conclusions of theorem 3.8 hold. We want to show $\{\mu_n \circ T_n\}_{n=1}^\infty$ satisfies

$$\sup_{n \in \mathbb{N}} \mathcal{C}_{\tilde{\epsilon}_n}(\mu_n \circ T_n; \tilde{\eta}) < \infty \qquad (19)$$

for a sequence $\tilde{\epsilon}_n > 0$ with $\tilde{\epsilon}_n \to 0$ and $\epsilon_n/\tilde{\epsilon}_n \to 1$ that will be chosen shortly. If so then by (18) there exists a subsequence $\mu_{n_m} \circ T_{n_m}$ and $\mu \in L^1(X; \{0, 1\})$ such that $\mu_{n_m} \circ T_{n_m} \to \mu$ in $L^1$ and therefore, $\mu_{n_m} \to \mu$ in $TL^1$. To show (19) we write

$$\mathcal{C}_{\tilde{\epsilon}_n}(\mu_n \circ T_n; \tilde{\eta}) = \frac{1}{\tilde{\epsilon}_n} \int_X V(\mu_n(T_n(x))) \rho(x) \, dx$$

$$+ \frac{1}{\tilde{\epsilon}_n^{d+1}} \int_{X^2} \tilde{\eta} \left( \frac{y-x}{\tilde{\epsilon}_n} \right) |\mu_n(T_n(x)) - \mu_n(T_n(y))| \, \rho(x) \rho(y) \, dy \, dx.$$

The first term is uniformly bounded, since by (16)

$$\frac{1}{\tilde{\epsilon}_n} \int_X V(\mu_n(T_n(x))) \rho(x) \, dx = \frac{\epsilon_n}{\tilde{\epsilon}_n} \frac{1}{\epsilon_n} \sum_{i=1}^n V(\mu_n(\xi_i)).$$

Assume that $|y - x/\tilde{\epsilon}_n| < b$ then

$$
\begin{aligned}
|T_n(x) - T_n(y)| &\leqslant |T_n(x) - x| + |x - y| + |y - T_n(y)| \\
&\leqslant 2\|\mathrm{Id} - T_n\|_{L^\infty(X)} + |x - y| \\
&\leqslant 2\|\mathrm{Id} - T_n\|_{L^\infty(X)} + b\tilde{\epsilon}_n.
\end{aligned}
$$

Choose $\tilde{\epsilon}_n$ satisfying

$$
2\|T_n - \mathrm{Id}\|_{L^\infty(X)} + b\tilde{\epsilon}_n = b\epsilon_n,
$$

that is, $\tilde{\epsilon}_n = \epsilon_n - ((2\|T_n - \mathrm{Id}\|_{L^\infty(X)})/(b))$. By the decay assumption on $\epsilon_n$ for $n$ sufficiently large (with probability one) $\tilde{\epsilon}_n > 0$, $\tilde{\epsilon}_n \to 0$ and $\tilde{\epsilon}_n/\epsilon_n \to 1$. Also

$$
\tilde{\eta}\left(\frac{x - y}{\tilde{\epsilon}_n}\right) = a \Rightarrow \tilde{\eta}\left(\frac{T_n(x) - T_n(y)}{\epsilon_n}\right) = a.
$$

Therefore,

$$
\tilde{\eta}_{\tilde{\epsilon}_n}(y - x) = \frac{1}{\tilde{\epsilon}_n^d}\tilde{\eta}\left(\frac{x - y}{\tilde{\epsilon}_n}\right) \leqslant \frac{1}{\tilde{\epsilon}_n^d}\tilde{\eta}\left(\frac{T_n(x) - T_n(y)}{\epsilon_n}\right) = \frac{\epsilon_n^d}{\tilde{\epsilon}_n^d}\tilde{\eta}_{\epsilon_n}(T_n(x) - T_n(y)).
$$

So,

$$
\begin{aligned}
\frac{1}{\tilde{\epsilon}_n}&\int_{X^2}\tilde{\eta}_{\tilde{\epsilon}_n}(y - x)\,|\mu_n(T_n(x)) - \mu_n(T_n(y))|\,\rho(x)\rho(y)\,\mathrm{d}y\,\mathrm{d}x \\
&\leqslant \frac{\epsilon_n^d}{\tilde{\epsilon}_n^{d+1}}\int_{X^2}\tilde{\eta}_{\epsilon_n}(T_n(x) - T_n(y))\,|\mu_n(T_n(x)) - \mu_n(T_n(y))|\,\rho(x)\rho(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \frac{\epsilon_n^d}{\tilde{\epsilon}_n^{d+1}}\sum_{i,j}\tilde{\eta}_{\epsilon_n}(\xi_i - \xi_j)\,|\mu_n(\xi_i) - \mu_n(\xi_j)| \quad \text{by (16)} \\
&= \frac{\epsilon_n^{d+1}}{\tilde{\epsilon}_n^{d+1}}\mathcal{E}_n(\mu_n;\tilde{\eta}) \\
&\leqslant \frac{\epsilon_n^{d+1}}{\tilde{\epsilon}_n^{d+1}}\mathcal{E}_n(\mu_n;\eta).
\end{aligned}
$$

It follows that the second term is also uniformly bounded in $n$.                    □

## 5. Γ-Convergence

The main result of this section is theorem 5.1, which states that $\mathcal{E}_n$ Γ-converge to $\mathcal{E}_\infty$.

THEOREM 5.1. *Under the same conditions as theorem 2.3*

$$
\mathcal{E}_\infty = \Gamma\text{-}\lim_{n\to\infty}\mathcal{E}_n
$$

*in the $TL^1$ sense and with probability one.*

The proof is a consequence of lemmas 5.2 and 5.4.

LEMMA 5.2 The lim inf inequality. *Under the same conditions as theorem* [2.3](#) *if* $\mu \in L^1(X)$ *and* $\mu_n \to \mu$ *in* $TL^1$ *then*

$$\mathcal{E}_\infty(\mu) \leqslant \liminf_{n \to \infty} \mathcal{E}_n(\mu_n)$$

*with probability one.*

*Proof.* Recall $\eta = \varphi \circ \pi$. Let $\mu_n \in L^1(\Psi_n), \mu \in L^1(X)$ with $\mu_n \to \mu$ in $TL^1$. Let $\nu_n = \mu_n \circ T_n \in L^1(X)$ where $T_n : X \to \Psi_n$ is as in theorem [3.8](#) (with probability one) so $\nu_n \to \mu$ in $L^1(X)$. Without loss of generality, we assume that

$$\liminf_{n \to \infty} \mathcal{E}_n(\mu_n) < \infty$$

else there is nothing to prove. By theorem [4.1](#) $\mu \in L^1(X; \{0,1\})$ hence the proof is complete if

$$\liminf_{n \to \infty} GTV_n(\mu_n) \geqslant TV(\mu; \rho, \eta) \tag{20}$$

where $GTV_n$ is defined by [(3)](#).

We show [(20)](#) in two steps.

**Step 1.** Assume $\rho$ is Lipschitz continuous on $X$.

**Step 2.** Generalize to continuous densities.

*Step 1.* Let $X'$ be a compact subset of $X$. We have

$$GTV_n(\mu_n) = \frac{1}{\epsilon_n} \frac{1}{n^2} \sum_{i,j} W_{ij} \left| \mu_n(\xi_i) - \mu_n(\xi_j) \right|$$

$$= \frac{1}{\epsilon_n^{d+1}} \int_{X^2} \eta \left( \frac{T_n(x) - T_n(y)}{\epsilon_n} \right)$$
$$\times \left| \nu_n(x) - \nu_n(y) \right| \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y \quad \text{using } (16)$$

$$\geqslant \frac{1}{\epsilon_n^{d+1}} \int_{X'} \int_X \eta \left( \frac{T_n(x) - T_n(y)}{\epsilon_n} \right) \left| \nu_n(x) - \nu_n(y) \right| \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y$$

$$= \frac{1}{\epsilon_n} \int_{X'} \int_{y+\epsilon_n z \in X} \eta \left( \frac{T_n(y + \epsilon_n z) - T_n(y)}{\epsilon_n} \right)$$
$$\times \left| \nu_n(y + \epsilon_n z) - \nu_n(y) \right| \rho(y + \epsilon_n z)\rho(y) \, \mathrm{d}z \, \mathrm{d}y$$

$$= \frac{1}{\epsilon_n} \int_{X'} \int_{y+\epsilon_n z \in X} \eta \left( \frac{T_n(y + \epsilon_n z) - T_n(y)}{\epsilon_n} \right)$$
$$\times \left| \nu_n(y + \epsilon_n z) - \nu_n(y) \right| \rho^2(y) \, \mathrm{d}z \, \mathrm{d}y + a_n$$

where

$$a_n = \frac{1}{\epsilon_n} \int_{X'} \int_{y+\epsilon_n z \in X} \eta \left( \frac{T_n(y + \epsilon_n z) - T_n(y)}{\epsilon_n} \right) \left| \nu_n(y + \epsilon_n z) \right.$$
$$\left. - \nu_n(y) \right| \rho(y) \left( \rho(y + \epsilon_n z) - \rho(y) \right) \, \mathrm{d}z \, \mathrm{d}y.$$

Assume that the support of $\varphi$ is contained within $B(0, M)$ then the following calculation shows that for any $y \in X$ that if $|z| \geqslant \tilde{M}$ where $\tilde{M} = M/\alpha + 2 \sup_{n \in \mathbb{N}}((\|T_n - \mathrm{Id}\|_{L^\infty})/(\epsilon_n)) < \infty$ then $\eta(((T_n(y + \epsilon_n z) - T_n(y))/(\epsilon_n))) = 0$:

$$M \leqslant \alpha \epsilon_n |z| - 2\alpha \frac{\|T_n - \mathrm{Id}\|_{L^\infty}}{\epsilon_n}$$

$$\leqslant \frac{\alpha}{\epsilon_n} \left( \epsilon_n |z| - |T_n(y + \epsilon_n z) - T_n(y) - \epsilon_n z| \right)$$

$$\leqslant \frac{\alpha}{\epsilon_n} \left| T_n(y + \epsilon_n z) - T_n(y) \right|$$

$$\leqslant \pi \left( \frac{T_n(y + \epsilon_n z) - T_n(y)}{\epsilon_n} \right).$$

It follows that

$$|a_n| \leqslant \frac{\mathrm{Lip}(\rho)\tilde{M}}{\inf_{x \in X} \rho(x)} \int_X \int_{y + \epsilon z \in X} \eta \left( \frac{T_n(y + \epsilon_n z) - T_n(y)}{\epsilon_n} \right)$$

$$\times |\nu_n(y + \epsilon_n z) - \nu_n(y)| \, \rho(y)\rho(y + \epsilon_n z) \, \mathrm{d}z \, \mathrm{d}y$$

$$= \frac{\tilde{M}\mathrm{Lip}(\rho)}{\inf_{x \in X} \rho(x)} \epsilon_n GTV_n(\mu_n)$$

$$\to 0 \quad \text{as } n \to \infty.$$

Let $\alpha_n$ and $c_n$ be as in definition 2.1 for $\delta = \frac{2\|T_n - \mathrm{Id}\|_{L^\infty}}{\epsilon_n}$ then since

$$\left| \frac{T_n(y + \epsilon_n z) - T_n(y)}{\epsilon_n} - z \right| \leqslant \frac{1}{\epsilon_n} \left( |T_n(y + \epsilon_n z) - y - \epsilon_n z| + |T_n(y) - y| \right)$$

$$\leqslant \frac{2\|T_n - \mathrm{Id}\|_{L^\infty}}{\epsilon_n}$$

we have that

$$\eta \left( \frac{T_n(y + \epsilon_n z) - T_n(y)}{\epsilon_n} \right) \geqslant c_n \eta \left( \alpha_n z \right).$$

So,

$$GTV_n(\mu_n) \geqslant \frac{c_n}{\epsilon_n} \int_{X'} \int_{y + \epsilon_n x \in X} \eta(\alpha_n z) \, |\nu_n(y + \epsilon_n z) - \nu_n(y)| \, \rho^2(y) \, \mathrm{d}z \, \mathrm{d}y + o(1)$$

$$= \frac{c_n}{\alpha_n^d \epsilon_n} \int_{X'} \int_{y + \epsilon_n \tilde{z}/\alpha_n \in X} \eta(\tilde{z}) \left| \nu_n \left( y + \frac{\epsilon_n \tilde{z}}{\alpha_n} \right) - \nu_n(y) \right| \rho^2(y) \, \mathrm{d}\tilde{z} \, \mathrm{d}y + o(1)$$

$$= \frac{c_n}{\alpha_n^d \epsilon_n} \int_{X'} \int_{\mathbb{R}^d} \eta(\tilde{z}) \left| \nu_n \left( y + \frac{\epsilon_n \tilde{z}}{\alpha_n} \right) - \nu_n(y) \right| \rho^2(y) \, \mathrm{d}\tilde{z} \, \mathrm{d}y - b_n + o(1)$$

where

$$b_n = \frac{c_n}{\alpha_n^d \epsilon_n} \int_{X'} \int_{y + \epsilon_n \tilde{z}/\alpha_n \notin X} \eta(\tilde{z}) \left| \nu_n \left( y + \frac{\epsilon_n \tilde{z}}{\alpha_n} \right) - \nu_n(y) \right| \rho^2(y) \, \mathrm{d}\tilde{z} \, \mathrm{d}y$$

By Fatou's lemma, proposition 3.4 and since $c_n, \alpha_n \to 1$ we have

$$
\liminf_{n \to \infty} \frac{c_n}{\alpha_n^d \epsilon_n} \int_{X'} \int_{\mathbb{R}^d} \eta(\tilde{z}) \left| \nu_n \left( y + \frac{\epsilon_n \tilde{z}}{\alpha_n} \right) - \nu_n(y) \right| \rho^2(y) \, \mathrm{d}\tilde{z} \, \mathrm{d}y
$$
$$
\geqslant \int_{\mathbb{R}^d} \eta(z) TV_z(\mu; \rho) \, \mathrm{d}z
$$
$$
= TV(\mu; \rho, \eta, X').
$$

Since $X'$ and $X^c$ are both closed and disjoint then $\tau := \mathrm{dist}(X', X^c) > 0$ and therefore, if $y \in X'$ and $y + \epsilon_n \tilde{z}/\alpha_n \notin X$ then $|\tilde{z}| \geqslant \alpha_n \tau/\epsilon_n$. Choose $n$ sufficiently large such that $\alpha_n \tau/\epsilon_n \geqslant M$ where the support of $\eta$ is contained in $B(0, M)$. Then for any pair $y, \tilde{z}$ satisfying the above condition we have $\eta(z) = \varphi \circ \pi(\tilde{z}) = 0$. Hence $b_n = 0$ for $n$ sufficiently large.

We have shown that

$$
\liminf_{n \to \infty} GTV_n(\mu_n) \geqslant TV(\mu; \rho, \varphi \circ \pi, X').
$$

If we consider a sequence $X'_m$ such that $X'_m \subset X'_{m+1}$ and $\mathbb{I}_{X'} \to \mathbb{I}_X$ pointwise then by the monotone convergence theorem $TV(\mu; \rho, \varphi \circ \pi, X'_m) \to TV(\mu; \rho, \varphi \circ \pi, X)$. This completes step 1.

*Step 2.* Denote the dependence of $\rho$ on $GTV_n$ by $GTV_n(\cdot; \rho)$. Assume $\rho : X \to [0, \infty)$ is continuous and let $\rho_k : \mathbb{R}^d \to [0, \infty)$ be defined by

$$
\rho_k(x) = \begin{cases} \inf_{y \in X} \left( \rho(y) + k|x - y| \right) & \text{if } x \in X \\ 0 & \text{otherwise.} \end{cases} \tag{21}
$$

Clearly $\rho_k(x) \leqslant \rho(x)$ for all $x \in X$. For any $y \in X \setminus B(x, \rho(x)/k)$ we have $|x - y| \geqslant \rho(x)/k$ and, therefore,

$$
\rho(y) + k|x - y| \geqslant \rho(y) + \rho(x) > \rho(x) \geqslant \rho_k(x).
$$

Then it follows that any (approximate) minimizer $y \in X$ of (21) must be contained in $B(x, \rho(x)/k)$. Hence $\rho_k(x) = \inf\{\rho(y) + k|x - y| : y \in B(x, \rho(x)/k)\}$ and, therefore,

$$
\rho(x) \geqslant \rho_k(x) \geqslant \inf \left\{ \rho(y) : y \in B\left( x, \frac{\rho(x)}{k} \right) \right\}.
$$

As $\rho$ is bounded above on $X$ then the previous inequality implies $\rho_k(x) \to \rho(x)$ for each $x \in X$. It is also clear that $\rho_k(x) \geqslant \inf_{x \in X} \rho(x) > 0$. Furthermore, for $x, z \in X$

$$
\rho_k(x) - \rho_k(z) = \inf_{y_1 \in X} \sup_{y_2 \in X} \rho(y_1) - \rho(y_2) + k\left(|x - y_1| - |z - y_2|\right)
$$
$$
\leqslant \sup_{y_2 \in X} k\left(|x - y_2| - |z - y_2|\right)
$$
$$
\leqslant k|x - z|
$$

so $\rho_k$ is Lipschitz in $X$. By step 1

$$\liminf_{n\to\infty} GTV_n(\mu_n; \rho) \geqslant \liminf_{n\to\infty} GTV_n(\mu_n; \rho_k) \geqslant TV(\mu; \rho_k, \eta).$$

By theorem 3.5

$$TV(\mu; \rho_k, \eta) = \int_X \rho_k^2(x)\sigma(x)\,\hat{\lambda}(\mathrm{d}x).$$

And therefore, by the monotone convergence theorem, one has

$$\lim_{k\to\infty} TV(\mu; \rho_k, \eta) = \int_X \rho(x)^2\sigma(x)\,\hat{\lambda}(\mathrm{d}x) = TV(\mu; \rho, \eta)$$

which completes the proof.                                                    □

For $\mu \notin L^1(X; \{0,1\})$ the recovery sequence is trivial as $\mathcal{E}_\infty(\mu) = \infty$. For $\mu \in L^1(X; \{0,1\})$ we show that it is enough to prove the existence of a recovery sequence when $\mu$ is a polyhedral function (defined below). Recall that $\mathcal{H}^k$ is the $k$-dimensional Hausdorff measure.

DEFINITION 5.3. *A (d-dimensional) polyhedral set in $\mathbb{R}^d$ is an open set $F$ whose boundary is a Lipschitz manifold contained in the union of finitely many affine hyperplanes. We say $\mu \in BV(X; \{0,1\})$ is a polyhedral function if there exists a polyhedral set $F$ such that $\partial F$ is transversal to $\partial X$ (i.e. $\mathcal{H}^{d-1}(\partial F \cup \partial X) = 0$) and $\mu(x) = 1$ for $x \in X \cap F$, $\mu(x) = 0$ for $x \in X \setminus F$.*

LEMMA 5.4 The existence of a recovery sequence for theorem 5.1. *Under the same conditions as theorem 2.3 for any $\mu \in L^1(X)$ there exists a sequence $\mu_n \to \mu$ in $TL^1$ such that*

$$\mathcal{E}_\infty(\mu) \geqslant \limsup_{n\to\infty} \mathcal{E}_n(\mu_n) \tag{22}$$

*with probability one.*

*Proof.* Without loss of generality assume $\mu \in BV(X; \rho, \eta) \cap L^1(X; \{0,1\})$. By the following (diagonalization) argument, it is enough to prove the lemma for polyhedral functions. Suppose the lemma holds for polyhedral functions and let $\mu \in BV(X; \rho, \eta) \cap L^1(X; \{0,1\})$. There exists a sequence of polyhedral functions $\mu_m \to \mu$ in $L^1$ and $TV(\mu_m; \rho, \eta) \to TV(\mu; \rho, \eta)$, for example, see [36, §9.4.1 lemma 1]. By passing to a subsequence, we may assume that

$$\|\mu_m - \mu\|_{L^1} \leqslant \tfrac{1}{m} \qquad \text{and} \qquad |\mathcal{E}_\infty(\mu_m) - \mathcal{E}_\infty(\mu)| \leqslant \tfrac{1}{m}.$$

Now for each $m$ there exists a sequence $\mu_n^{(m)}$ and a limit $\mu^{(m)}$ (as $n \to \infty$ for each $m$) such that

$$\mu_n^{(m)} \to \mu^{(m)} \quad \text{in } L^1 \text{ as } n \to \infty \qquad \text{and} \qquad \mathcal{E}_\infty(\mu^{(m)}) \geqslant \limsup_{n\to\infty} \mathcal{E}_n(\mu_n^{(m)}).$$

For each $m$ there exists $N_m$ such that

$$\left\|\mu_n^{(m)} - \mu^{(m)}\right\|_{L^1} \leqslant \frac{1}{m} \qquad \text{and} \qquad \mathcal{E}_\infty(\mu^{(m)}) \geqslant \mathcal{E}_n(\mu_n^{(m)}) - \frac{1}{m}$$

for all $n \geqslant N_m$. Let $\mu_m = \mu_{N_m}^{(m)}$ then we have

$$\|\mu_m - \mu\|_{L^1} \leqslant \left\|\mu_{N_m}^{(m)} - \mu^{(m)}\right\|_{L^1} + \left\|\mu^{(m)} - \mu\right\|_{L^1} \leqslant \frac{2}{m} \qquad \text{and}$$

$$\mathcal{E}_\infty(\mu) \geqslant \mathcal{E}_\infty(\mu^{(m)}) - \frac{1}{m} \geqslant \mathcal{E}_n(\mu_m) - \frac{2}{m}.$$

Hence it is enough to prove the theorem for polyhedral functions.

Therefore, assume $\mu \in BV(X; \{0,1\})$ is a polyhedral function corresponding to the polyhedral set $F$, that is, $\mu = \mathbb{I}_F$. Let $\mu_n$ be the restriction of $\mu$ to $\Psi_n$. Define $T_n$ as in theorem 3.8 and use it to create a partition of $X$, $X = \cup_{i=1}^n T_n^{-1}(\xi_i)$. If $x, y \in T_n^{-1}(\xi_i)$ then

$$|x - y| \leqslant |x - T_n(x)| + |T_n(x) - T_n(y)| + |T_n(y) - y| \leqslant 2\|\mathrm{Id} - T_n\|_{L^\infty(X)}.$$

Let $x \in T_n^{-1}(\xi_i)$ and assume $\mathrm{dist}(\partial F, x) > 2\|\mathrm{Id} - T_n\|_{L^\infty}$. If $y \in T_n^{-1}(\xi_i)$ then $\mu(y) = \mu(x)$ (since $T_n^{-1}(\xi_i) \subset B(x, 2\|\mathrm{Id} - T_n\|_{L^\infty(X)})$ and $B(x, 2\|\mathrm{Id} - T_n\|_{L^\infty(X)}) \cap \partial F = \emptyset$). Therefore,

$$\int_{T_n^{-1}(\xi_i)} |\mu_n(T_n(y)) - \mu(y)|\, \rho(y)\, \mathrm{d}y = 0.$$

In particular,

$$\int_X |\mu_n(T_n(y)) - \mu(y)|\, \rho(y)\, \mathrm{d}y = \int_{X_n} |\mu_n(T_n(y)) - \mu(y)|\, \rho(y)\, \mathrm{d}y$$
$$\leqslant \|\rho\|_{L^\infty(X)} \mathrm{Vol}(X_n)$$

where

$$X_n = \left\{ y \in X \; : \; \mathrm{dist}(\partial F, y) \leqslant 2\|\mathrm{Id} - T_n\|_{L^\infty(X)} \right\}.$$

Clearly, $\mathrm{Vol}(X_n) = O(\|\mathrm{Id} - T_n\|_{L^\infty(X)}) = o(1)$ and therefore, $\mu_n \to \mu$ in $TL^1$. Define $\nu_n = \mu_n \circ T_n$ then since $\nu_n, \mu \in L^1(X; \{0,1\})$ we have that (22) is equivalent to

$$TV(\mu; \rho, \eta) \geqslant \limsup_{n \to \infty} GTV_n(\mu_n).$$

We complete the proof in two steps.

**Step 1.** Assume $\rho$ is Lipschitz on $X$.

**Step 2.** Generalize to continuous densities.

*Step 1.* We can write assumption 3 in definition 2.1 as for any $\delta > 0$ there exists $c_\delta, \alpha_\delta$ such that if $|x - y| < \delta$ then $\eta(x) \leqslant 1/c_\delta \eta(w/\alpha_\delta)$ and $c_\delta, \alpha_\delta \to 1$ as $\delta \to 0$.

Let $c_n, \alpha_n$ be such constants for $\delta = ((2\|T_n - \mathrm{Id}\|_{L^\infty})/(\epsilon_n))$ since

$$\left| \frac{T_n(x) - T_n(y)}{\epsilon_n} - \frac{x-y}{\epsilon_n} \right| \leqslant 2 \frac{2\|T_n - \mathrm{Id}\|_{L^\infty}}{\epsilon_n} = \delta$$

then

$$\eta\left( \frac{T_n(x) - T_n(y)}{\epsilon_n} \right) \leqslant \frac{1}{c_n} \eta\left( \frac{x-y}{\tilde{\epsilon}_n} \right)$$

where $\tilde{\epsilon} = \epsilon_n \alpha_n$.

So,

$$\begin{aligned}
GTV_n(\mu_n) &= \frac{1}{\epsilon_n^{d+1}} \int_{X^2} \eta\left( \frac{T_n(x) - T_n(y)}{\epsilon} \right) |\nu_n(x) - \nu_n(y)| \, \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y \\
&\leqslant \frac{\alpha_n^{d+1}}{c_n \tilde{\epsilon}_n^{d+1}} \int_{X^2} \eta\left( \frac{x-y}{\tilde{\epsilon}_n} \right) |\nu_n(x) - \nu_n(y)| \, \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \frac{\alpha_n^{d+1}}{c_n} CTV_n(\nu_n)
\end{aligned}$$

with $CTV_n$ defined below. Let us approximate $\mu$ by a sequence $\zeta_n \in C^\infty(\mathbb{R}^d) \cap BV(X)$ such that $\zeta_n \to \mu$ in $L^1(X)$ and $TV(\zeta_n; \rho, \eta) \to TV(\mu; \rho, \eta)$. Without loss of generality assume that $\zeta_n(x) = 0$ for all $x \in \mathbb{R}^d \setminus X$ and $\|\mu - \zeta_n\|_{L^1(X)} = o(\tilde{\epsilon}_n)$ (by recourse to a subsequence of $\zeta_n$ and relabelling). Then

$$\begin{aligned}
&CTV_n(\zeta_n) \\
&:= \frac{1}{\tilde{\epsilon}_n^{d+1}} \int_{X^2} \eta\left( \frac{x-y}{\tilde{\epsilon}_n} \right) |\zeta_n(x) - \zeta_n(y)| \, \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y \\
&= \frac{1}{\tilde{\epsilon}_n^{d+1}} \int_{X^2} \eta\left( \frac{x-y}{\tilde{\epsilon}_n} \right) \left| \int_0^1 \nabla\zeta_n(y + s(x-y)) \cdot (x-y) \, \mathrm{d}s \right| \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}y \\
&\leqslant \frac{1}{\tilde{\epsilon}_n^{d+1}} \int_X \int_0^1 \int_X \eta\left( \frac{x-y}{\tilde{\epsilon}_n} \right) |\nabla\zeta_n(y + s(x-y)) \cdot (x-y)| \, \rho(x)\rho(y) \, \mathrm{d}x \, \mathrm{d}s \, \mathrm{d}y \\
&= \int_X \int_0^1 \int_{\mathcal{Z}_{nhs}} \eta(z) \, |\nabla\zeta_n(h) \cdot z| \, \rho(h + (1-s)\tilde{\epsilon}_n z)\rho(h - \tilde{\epsilon}_n s z) \, \mathrm{d}z \, \mathrm{d}s \, \mathrm{d}h \\
&\leqslant TV(\zeta_n; \rho, \eta) + c_n
\end{aligned}$$

where

$$\mathcal{Z}_{nhs} = \left\{ z \in \mathbb{R}^d : h + (1-s)\tilde{\epsilon}_n z \in X \text{ and } h - \tilde{\epsilon}_n s z \in X \right\}$$

$$\begin{aligned}
c_n &= \int_X \int_0^1 \int_{\mathcal{Z}_{nhs}} \eta(z) \, |\nabla\zeta_n(h) \cdot z| \\
&\quad \times \left( \rho(h + (1-s)\tilde{\epsilon}_n z)\rho(h - \tilde{\epsilon}_n s z) - \rho^2(h) \right) \mathrm{d}z \, \mathrm{d}s \, \mathrm{d}h.
\end{aligned}$$

If

$$\lim_{n\to\infty} c_n = 0 \tag{23}$$

$$\text{and} \quad \lim_{n\to\infty} |CTV_n(\nu_n) - CTV_n(\zeta_n)| = 0 \tag{24}$$

then

$$\limsup_{n\to\infty} CTV_n(\nu_n) = \limsup_{n\to\infty} CTV_n(\zeta_n) \leqslant \limsup_{n\to\infty} TV(\zeta_n; \rho, \eta) = TV(\mu; \rho, \eta).$$

We now show (23). Since the support of $\eta$ is bounded then there exists $M > 0$ such that $\mathrm{spt}(\eta) \subset B(0, M)$. For any $|z| > M$, we have $\eta(z) = 0$ and for any $|z| \leqslant M$ one can show

$$\left| \rho^2(h) - \rho(h + (1-s)\tilde{\epsilon}_n z)\rho(h + \tilde{\epsilon}_n sz) \right| \leqslant \|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) \left( |(1-s)\tilde{\epsilon}_n z| + |\tilde{\epsilon}_n sz| \right)$$

$$\leqslant 2\tilde{\epsilon}_n \|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) M.$$

Then

$$|c_n| \leqslant 2\|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) \tilde{\epsilon}_n M \int_X \int_{\mathbb{R}^d} \eta(z) |\nabla \zeta_n(h) \cdot z| \, \mathrm{d}z \, \mathrm{d}h$$

$$\leqslant \frac{2\|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) \tilde{\epsilon}_n M}{\inf_{x \in X} \rho^2(x)} \int_X \int_{\mathbb{R}^d} \eta(z) |\nabla \zeta_n(h) \cdot z| \, \rho^2(h) \, \mathrm{d}z \, \mathrm{d}h$$

$$= \frac{2\|\rho\|_{L^\infty(X)} \mathrm{Lip}(\rho) \tilde{\epsilon}_n M}{\inf_{x \in X} \rho^2(x)} TV(\zeta_n; \rho, \eta).$$

To complete step 1, we show (24). This follows by:

$$|CTV_n(\nu_n) - CTV_n(\zeta_n)|$$

$$\leqslant \frac{\|\rho\|_{L^\infty(X)}^2}{\tilde{\epsilon}_n^{d+1}} \int_{X^2} \eta\left(\frac{y-x}{\tilde{\epsilon}_n}\right) \left( |\nu_n(x) - \zeta_n(x)| + |\nu_n(y) - \zeta_n(y)| \right) \mathrm{d}x \, \mathrm{d}y$$

$$\leqslant \frac{2\|\varphi\|_{L^\infty(\mathbb{R}^d)} \mathrm{Vol}(X) \|\rho\|_{L^\infty(X)}^2}{\tilde{\epsilon}_n} \int_X |\nu_n(x) - \zeta_n(x)| \, \mathrm{d}y$$

$$\leqslant \frac{2\|\varphi\|_{L^\infty(\mathbb{R}^d)} \mathrm{Vol}(X) \|\rho\|_{L^\infty(X)}^2}{\tilde{\epsilon}_n} \left( \|\nu_n - \mu\|_{L^1(X)} + \|\mu - \zeta_n\|_{L^1(X)} \right)$$

$$\to 0$$

where the last line follows as $\|\mu - \zeta_n\|_{L^1(X)} = o(\tilde{\epsilon}_n)$ and $\|\nu_n - \mu\|_{L^1(X)} = O(\|T_n - \mathrm{Id}\|_{L^\infty(X)}) = o(\epsilon_n) = o(\tilde{\epsilon}_n)$.

*Step 2.* Let $GTV(\cdot; \rho)$ be the graph total variation defined using $\rho$. Let $\rho$ be continuous but not necessarily Lipschitz and define $\rho_k : \mathbb{R}^d \to [0, \infty)$ by

$$\rho_k(x) = \begin{cases} \sup_{y \in X} \rho(y) - k|x - y| & \text{if } x \in X \\ 0 & \text{otherwise.} \end{cases}$$

Similarly, to lemma 5.2 step 2, we can check that $\rho_k$ is bounded above and below by positive constants, Lipschitz continuous on $X$ and converges pointwise to $\rho$ from

above. We have

$$\limsup_{n\to\infty} GTV_n(\mu_n; \rho) \leqslant \limsup_{n\to\infty} GTV_n(\mu_n; \rho_k) \leqslant TV(\mu; \rho_k, \eta).$$

By the monotone convergence theorem and theorem 3.5, we have $\lim_{k\to\infty} TV(\mu; \rho_k, \eta) = TV(\mu; \rho, \eta)$.                                                            □

## 6. Preliminary results for the rate of convergence

There are two main sources of error defined to be $|\mathcal{E}_n(\mu) - \mathcal{E}_\infty(\mu)|$. The first is due to statistical fluctuations, for example, data points that lie in the tails that have a large impact when $n$ is small. The second is due to systematic bias. Systematic bias is due to several factors such as the details of the scaling and the geometry of the problem. In this section, by taking the expectation over the data, we only consider the second source of error. In this section, we will restrict ourselves to when $\mu = \mathbb{I}_E$ is a polyhedral function (see definition 5.3).

We first fix our notation. The boundary of a polyhedral set $E$ is contained in the union of finitely many ($N$ say) affine hyperplanes $H_i$ for $i = 1, \ldots, N$. We call the set $\partial E_i := \partial E \cap H_i$ a face of $E$. By construction $\partial E = \cup_{i=1}^N \partial E_i$. The intersection of two faces is an edge $e_{ij} = \partial E_i \cap \partial E_j$. It is unfortunate but unavoidable that we use the term edge in two different contexts; either as an edge in a graph, or as an edge of a polyhedral set. It should be clear from the context what is meant. The intersection of edges is called a corner.

To reduce bias let us redefine the normalization on $GTV_n$ so that

$$GTV_n(\mu) = \frac{1}{\epsilon_n} \frac{1}{n(n-1)} \sum_{i,j} \eta_{\epsilon_n}(\xi_i - \xi_j) |\mu(\xi_i) - \mu(\xi_j)|. \tag{25}$$

For $\Xi_{ij} = 1/\epsilon_n \eta_{\epsilon_n}(\xi_i - \xi_j) |\mu(\xi_i) - \mu(\xi_j)|$ one can write

$$GTV_n(\mu) = \frac{1}{n(n-1)} \sum_{i,j} \Xi_{ij}.$$

Since $\xi_i$ are iid then

$$\mathbb{E}\Xi_{ij} = \begin{cases} \mathbb{E}\Xi_{12} & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

Hence $\mathbb{E}GTV_n(\mu) = \mathbb{E}\Xi_{12}$. This would not be true for the normalization we considered previously.

For simplicity, we make the following assumptions. Assume $X = (0,1)^d$ where $d \geqslant 1$ and that $\rho \equiv 1$ on $X$. We use an isotropic interaction potential $\eta = \mathbb{I}_{B(0,1)}$. These assumptions simplify the calculations which allows one to have a better understanding of the methodology without the notational burden if one used more general assumptions. We expect that the results in this section can be generalized to a wider class of interaction potentials $\eta$, spaces $X \subset \mathbb{R}^d$ and probability densities $\rho$. We start with the convergence of the expectation.

THEOREM 6.1. *Let $X = (0,1)^d$ with $d \geqslant 2$, $\rho \equiv 1$ and $\epsilon_n$ be any sequence converging to zero. The data are distributed $\xi_i \overset{\text{iid}}{\sim} \rho$ and let $\Psi_n = \{\xi_i\}_{i=1}^n$. Define $GTV_n : L^1(\Psi_n) \to [0,\infty]$ by (25) where the weights are given by $W_{ij} = \eta_{\epsilon_n}(\xi_i - \xi_j)$ and $\eta_{\epsilon_n}(x) = 1/\epsilon_n^d \mathbb{I}_{|x| \leqslant \epsilon_n}$. Define $TV(\cdot; \rho, \eta) : L^1(X) \to [0,\infty]$ by (8–10). Let $\mu = \mathbb{I}_E$ be a polyhedral function. Then*

$$|\mathbb{E} GTV_n(\mu) - TV(\mu; \rho, \eta)| = O(\epsilon_n).$$

Note that we do not need a lower bound on the decay of $\epsilon_n$. By taking the expectation we are immediately in the continuous setting and therefore, lose all the graphical structure. In particular,

$$\mathbb{E} GTV_n(\mu) = \frac{1}{\epsilon_n} \int_{(0,1)^d} \int_{(0,1)^d} \eta_{\epsilon_n}(x - y) |\mu(x) - \mu(y)| \, \mathrm{d}y \, \mathrm{d}x$$

has no discrete structure.

Our proof shows that along faces of $E$ and sufficiently far from edges in some sense the expected graph total variation is equal to the total variation. To be more precise if $E = \{x \in X : w \cdot x > 0\}$ for some $w \in \mathbb{R}^d$ then $E$ has no edges and therefore $GTV_n(\mu) = TV(\mu; \rho, \eta)$. The discrepancy between $\mathbb{E} GTV_n(\mu)$ and $TV(\mu; \rho, \eta)$, which we show is of order $\epsilon_n$, is a consequence of having to approximate along edges of $E$.

*Proof of Theorem 6.1.* Let $\partial E = \cup_{i=1}^N \partial E_i$. We first calculate $TV(\mu; \rho, \eta)$,

$$TV(\mu; \rho, \eta) = \int_{\partial \{\mu = 1\}} \sigma(n(x)) \, \mathrm{d}\mathcal{H}^{d-1}(x) = \sum_{i=1}^N |\partial E_i|_{\mathcal{H}^{d-1}} \, \sigma(n_i)$$

where $n_i$ is the outward unit normal for side $\partial E_i$ and we use $|\cdot|_{\mathcal{H}^{d-1}}$ to denote the $\mathcal{H}^{d-1}$ measure. Observe

$$\sigma = \sigma(n_i) = \int_{B(0,1)} |x \cdot n_i| \, \mathrm{d}x = \int_{B(0,1)} |x_d| \, \mathrm{d}x.$$

So $TV(\mu; \rho, \eta) = \sigma |\partial E|_{\mathcal{H}^{d-1}}$.

Consider the face $\partial E_i$, we will approximate this with a smaller face $\partial E_i^{(n)} \subset \partial E_i$ that will approximate the graph total variation to within $O(\epsilon_n)$. Consider the set,

$$R_i(s) = \Big\{ x + \epsilon_n t n : t \in [-1, 1], n \text{ is normal to } \partial E_i \text{ and}$$

$$x \in \partial E_i \text{ with } \mathrm{dist}(\partial(\partial E_i), x) > s \Big\}$$

where $\partial(\partial E_i)$ is the $d - $ two-dimensional boundary of the face $\partial E_i$. There exists $s_n = O(\epsilon_n)$ such that for all $i$

$$R_i(s_n) \cap \left( \bigcup_{j \neq i} \partial E_j \right) = \emptyset.$$

Define

$$B_n = \left\{ x \,:\, \mathrm{dist}(x, \partial E) < \epsilon_n \right\}$$

$$S_i^{(n)} = R_i(s_n + \epsilon_n) \qquad\qquad T_i^{(n)} = R_i(s_n)$$

$$U_i^{(n)} = \left\{ x + \epsilon_n t n \,:\, t \in [-1, 1], n \text{ is normal to } \partial E_i \text{ and} \right.$$

$$\left. x \in \partial E_i \text{ with } \mathrm{dist}(\partial(\partial E_i), x) \leqslant s_n + \epsilon_n \right\}$$

$$\partial E_i^{(n)} = \left\{ x \in \partial E_i \,:\, \mathrm{dist}(x, \partial(\partial E_i)) > s_n + \epsilon_n \right\}.$$

By construction $\bigcup_{x \in S_i^{(n)} \cap E} [E^c \cap B(x, \epsilon_n)] \subseteq (T_i^{(n)} \cap E^c)$ and $\bigcup_{x \in S_i^{(n)} \cap E^c}$ $[E \cap B(x, \epsilon_n)] \subseteq (T_i^{(n)} \cap E)$. Now,

$$\mathbb{E}GTV_n(\mu) = \frac{1}{\epsilon_n} \int_{B_n} \int_{B_n} \eta_{\epsilon_n}(x - y) \, |\mu(x) - \mu(y)| \, \mathrm{d}y \, \mathrm{d}x$$

$$= \frac{1}{\epsilon_n^{d+1}} \sum_{i=1}^{N} \int_{S_i^{(n)} \cap E} \int_{T_i^{(n)} \cap E^c} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x$$

$$+ \frac{1}{\epsilon_n^{d+1}} \int_{(B_n \cap E) \backslash \bigcup_{i=1}^{N} S_i^{(n)}} \int_{E^c} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x$$

$$+ \frac{1}{\epsilon_n^{d+1}} \sum_{i=1}^{N} \int_{S_i^{(n)} \cap E^c} \int_{T_i^{(n)} \cap E} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x$$

$$+ \frac{1}{\epsilon_n^{d+1}} \int_{(B_n \cap E^c) \backslash \bigcup_{i=1}^{N} S_i^{(n)}} \int_{E} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x.$$

We have

$$\frac{1}{\epsilon_n^{d+1}} \int_{(B_n \cap E) \backslash \bigcup_{i=1}^{N} S_i^{(n)}} \int_{E^c} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x \leqslant \frac{1}{\epsilon_n} \sum_{i=1}^{N} \int_{U_i^{(n)}} \int_{x + \epsilon_n z \in E^c} \mathbb{I}_{|z| \leqslant 1} \, \mathrm{d}z \, \mathrm{d}x$$

$$\leqslant \frac{1}{\epsilon_n} \sum_{i=1}^{N} \mathrm{Vol}(U_i^{(n)}) \mathrm{Vol}(B(0,1))$$

$$= O(\epsilon_n)$$

since $\mathrm{Vol}(U_i^{(n)}) = O(\epsilon_n(s_n + \epsilon_n)) = O(\epsilon_n^2)$. Similarly,

$$\frac{1}{\epsilon_n^{d+1}} \int_{(B_n \cap E^c) \backslash \bigcup_{i=1}^{N} S_i^{(n)}} \int_{E} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x = O(\epsilon_n).$$

For the remaining terms in the above expansion of $\mathbb{E}GTV_n(\mu)$, we consider the $i^{\mathrm{th}}$ face. After a suitable change of coordinates, we can assume that $\partial E_i \subset \{x \,:\, x_1 = 0\}$

and $n_i = (1, 0, \ldots, 0)$. Then

$$\frac{1}{\epsilon_n} \int_{S_i^{(n)} \cap E} \int_{T_i^{(n)} \cap E^c} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x$$

$$= \frac{1}{\epsilon_n} \int_{\partial E_i^{(n)}} \int_{-\epsilon_n}^0 \int_{y_1 > 0} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x_1 \, \mathrm{d}x_{2:d}$$

$$= \frac{1}{\epsilon_n} \int_{\partial E_i^{(n)}} \int_{-\epsilon_n}^0 \int_{x_1 = \epsilon_n z_1 > 0} \mathbb{I}_{|z| \leqslant 1} \, \mathrm{d}z \, \mathrm{d}x_1 \, \mathrm{d}x_{2:d}$$

$$= \frac{1}{\epsilon_n} \int_{\partial E_i^{(n)}} \int_{-\epsilon_n}^0 \int_{|z| \leqslant 1} \mathbb{I}_{x_1 + \epsilon_n z_1 > 0} \, \mathrm{d}z \, \mathrm{d}x_1 \, \mathrm{d}x_{2:d}$$

$$= \int_{\partial E_i^{(n)}} \int_{|z| \leqslant 1} z_1 \mathbb{I}_{z_1 > 0} \, \mathrm{d}z \, \mathrm{d}x_{2:d-1}$$

$$= \frac{1}{2} |\partial E_i^{(n)}|_{\mathcal{H}^{d-1}} \int_{|z| \leqslant 1} |z_1| \, \mathrm{d}z$$

$$= \frac{\sigma}{2} |\partial E_i^{(n)}|_{\mathcal{H}^{d-1}}$$

$$= \frac{\sigma}{2} |\partial E_i|_{\mathcal{H}^{d-1}} + O(\epsilon_n).$$

Analogously,

$$\frac{1}{\epsilon_n^{d+1}} \sum_{i=1}^N \int_{S_i^{(n)} \cap E^c} \int_{T_i^{(n)} \cap E} \mathbb{I}_{|x-y| \leqslant \epsilon_n} \, \mathrm{d}y \, \mathrm{d}x = \frac{\sigma}{2} |\partial E_i|_{\mathcal{H}^{d-1}} + O(\epsilon_n).$$

Collecting terms, we have shown

$$\mathbb{E} GTV_n(\mu) = TV(\mu; \rho, \eta) + O(\epsilon_n)$$

which completes the proof. $\qquad\square$

The above theorem established convergence in the mean of $GTV_n(\mu^n)$ to $TV(\mu; \rho, \eta)$. A natural next step is to establish convergence in mean square. The next theorem gives the asymptotic expansion of $\mathbb{E} |GTV_n(\mu) - TV(\mu; \rho, \eta)|^2$. We order the expansion so that dominant terms come first (the ordering follows from the scaling given by assumption 3 in definition 2.1). The dominant term depends on the bias $\alpha_n = \mathbb{E} GTV_n(\mu) - TV(\mu; \rho, \eta)$ which is $O(\epsilon_n)$ due to the approximation along the edges of each face of $E$. The complexity of refining this approximation, that is, finding the constant $c$ such that $\alpha_n = c\epsilon_n + \text{h.o.t.}$ goes beyond the scope of the paper.

THEOREM 6.2. *Under the same conditions as theorem* 6.1

$$\mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2$$

$$= -2\alpha_n TV(\mu;\rho,\eta) + \frac{4(n-2)|\partial E|_{\mathcal{H}^{d-1}}V}{n(n-1)\epsilon_n} + \frac{2TV(\mu;\rho,\eta)}{n(n-1)\epsilon_n^{d+1}}$$

$$+ \frac{(n-2)(n-3)\alpha_n^2}{n(n-1)} + \frac{(6-4n)TV(\mu;\rho,\eta)^2}{n(n-1)} + O\left(\frac{1}{n}\right)$$

$$+ O\left(\frac{1}{\epsilon_n^d n^2}\right) + \frac{2\alpha_n TV(\mu;\rho,\eta)}{n(n-1)}$$

*where* $\alpha_n = \mathbb{E}GTV_n(\mu) - TV(\mu;\rho,\eta) = O(\epsilon_n)$ *is the bias and*

$$V = \frac{1}{2}\int_{B(0,1)}\int_{B(0,1)} \min\{|z_d|,|y_d|\}\,\mathrm{d}z\,\mathrm{d}y.$$

*Proof.* We can write

$$\mathbb{E}\left|GTV_n(\mu) - TV(\mu;\rho,\eta)\right|^2$$

$$= \mathbb{E}GTV_n(\mu)^2 + TV(\mu;\rho,\eta)^2 - 2TV(\mu;\rho,\eta)\mathbb{E}GTV_n(\mu)$$

$$= \mathbb{E}GTV_n(\mu)^2 - TV(\mu;\rho,\eta)^2 - 2TV(\mu;\rho,\eta)\alpha_n.$$

Let $\Xi_{ij} = 1/\epsilon_n\eta_{\epsilon_n}(\xi_i - \xi_j)\left|\mu(\xi_i) - \mu(\xi_j)\right|$ then

$$GTV_n(\mu) = \frac{1}{n(n-1)}\sum_{i,j}\Xi_{ij} \quad \text{and} \quad GTV_n(\mu)^2 = \frac{1}{n^2(n-1)^2}\sum_{i,j,k,l}\Xi_{ij}\Xi_{kl}.$$

Let $i,j,k,l$ be distinct, then $GTV_n(\mu)^2$ has the following contributions:

A. $2n(n-1)$ terms consisting of $\Xi_{ij}^2$,

B. $4n(n-1)(n-2)$ terms consisting of $\Xi_{ij}\Xi_{ik}$ and

C. $n(n-1)(n-2)(n-3)$ terms consisting of $\Xi_{ij}\Xi_{kl}$.

For C we use independence of $\Xi_{ij}$ with $\Xi_{kl}$ to write

$$\mathbb{E}\Xi_{ij}\Xi_{kl} = \mathbb{E}\Xi_{ij}\mathbb{E}\Xi_{kl} = \left(\mathbb{E}GTV_n(\mu)\right)^2 = \left(TV(\mu;\rho,\eta) + \alpha_n\right)^2$$

$$= TV(\mu;\rho,\eta)^2 + 2\alpha_n TV(\mu;\rho,\eta) + \alpha_n^2.$$

For A:

$$\mathbb{E}\Xi_{ij}^2 = \frac{1}{\epsilon_n^{2d+2}}\int_{(0,1)^d}\int_{(0,1)^d} \mathbb{I}_{|x-y|\leqslant\epsilon_n}\left|\mu(x) - \mu(y)\right|\,\mathrm{d}x\,\mathrm{d}y$$

$$= \frac{1}{\epsilon_n^{d+1}}\mathbb{E}GTV_n(\mu)$$

$$= \frac{1}{\epsilon_n^{d+1}}TV(\mu;\rho,\eta) + O\left(\frac{1}{\epsilon_n^d}\right) \quad \text{by theorem } 6.1.$$

Now we consider B. We have

$$
\mathbb{E}\Xi_{ij}\Xi_{ik} = \frac{1}{\epsilon_n^{2d+2}} \int_{(0,1)^d} \int_{(0,1)^d} \int_{(0,1)^d} \mathbb{I}_{|x-y|\leqslant\epsilon_n} \mathbb{I}_{|x-w|\leqslant\epsilon_n} |\mu(x)-\mu(y)|
$$
$$
\times |\mu(x)-\mu(w)| \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
= \frac{1}{\epsilon_n^{2d+2}} \sum_{i=1}^{N} \int_{S_i^{(n)}\cap E} \int_{T_i^{(n)}\cap E^c} \int_{T_i^{(n)}\cap E^c} \mathbb{I}_{|x-y|\leqslant\epsilon_n} \mathbb{I}_{|x-w|\leqslant\epsilon_n} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
+ \frac{1}{\epsilon_n^{2d+2}} \int_{(B_n\cap E)\setminus\cup_{i=1}^{N}S_i^{(n)}} \int_{E^c} \int_{E^c} \mathbb{I}_{|x-y|\leqslant\epsilon_n} \mathbb{I}_{|x-w|\leqslant\epsilon_n} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
+ \frac{1}{\epsilon_n^{2d+2}} \sum_{i=1}^{N} \int_{S_i^{(n)}\cap E^c} \int_{T_i^{(n)}\cap E} \int_{T_i^{(n)}\cap E} \mathbb{I}_{|x-y|\leqslant\epsilon_n} \mathbb{I}_{|x-w|\leqslant\epsilon_n} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
+ \frac{1}{\epsilon_n^{2d+2}} \int_{(B_n\cap E^c)\setminus\cup_{i=1}^{N}S_i^{(n)}} \int_{E} \int_{E} \mathbb{I}_{|x-y|\leqslant\epsilon_n} \mathbb{I}_{|x-w|\leqslant\epsilon_n} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x.
$$

Now,

$$
\frac{1}{\epsilon_n^{2d+2}} \int_{(B_n\cap E)\setminus\cup_{i=1}^{N}S_i^{(n)}} \int_{E^c} \int_{E^c} \mathbb{I}_{|x-y|\leqslant\epsilon_n} \mathbb{I}_{|x-w|\leqslant\epsilon_n} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
= \frac{1}{\epsilon_n^{2}} \int_{(B_n\cap E)\setminus\cup_{i=1}^{N}S_i^{(n)}} \int_{x+\epsilon_n y\in E^c} \int_{x+\epsilon_n w\in E^c} \mathbb{I}_{|z|\leqslant1} \mathbb{I}_{|w|\leqslant1} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
\leqslant \frac{1}{\epsilon_n^{2}} \left(\mathrm{Vol}(B(0,1))\right)^2 \mathrm{Vol}((B_n\cap E)\setminus\cup_{i=1}^{N}S_i^{(n)})
$$
$$
= O(1).
$$

where $B_n, S_i^{(n)}$ and $T_i^{(n)}$ are as in the proof of Theorem 6.1. Considering each face individually, after rotating,

$$
\frac{1}{\epsilon_n^{2d+2}} \int_{S_i^{(n)}\cap E} \int_{T_i^{(n)}\cap E^c} \int_{T_i^{(n)}\cap E^c} \mathbb{I}_{|x-y|\leqslant\epsilon_n} \mathbb{I}_{|x-w|\leqslant\epsilon_n} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
= \frac{1}{\epsilon_n^{2}} \int_{S_i^{(n)}\cap E} \int_{x_d+\epsilon_n y_d>0} \int_{x_d+\epsilon_n w>0} \mathbb{I}_{|y|\leqslant1} \mathbb{I}_{|w|\leqslant1} \, \mathrm{d}w \, \mathrm{d}y \, \mathrm{d}x
$$
$$
= \frac{1}{\epsilon_n^{2}} \int_{\partial E_i^{(n)}} \int_{-\epsilon_n}^{0} \int_{|y|\leqslant1} \int_{|w|\leqslant1} \mathbb{I}_{x_d+\epsilon_n y_d>0} \mathbb{I}_{x_d+\epsilon_n w_d>0} \, \mathrm{d}y \, \mathrm{d}x_d \, \mathrm{d}x_{1:d-1}
$$
$$
= \frac{1}{\epsilon_n^{2}} \int_{\partial E_i^{(n)}} \int_{-\epsilon_n}^{0} \int_{|w|\leqslant1} \int_{|y|\leqslant1} \mathbb{I}_{x_d>-\epsilon_n\min\{w_d,y_d\}} \, \mathrm{d}y \, \mathrm{d}w \, \mathrm{d}x_d \, \mathrm{d}x_{1:d-1}
$$
$$
= \frac{1}{\epsilon_n} \int_{\partial E_i^{(n)}} \int_{|w|\leqslant1} \int_{|y|\leqslant1} \mathbb{I}_{\min\{w_d,y_d\}>0} \min\{w_d,y_d\} \, \mathrm{d}y \, \mathrm{d}w \, \mathrm{d}x_{1:d-1}
$$

$$= \frac{1}{4\epsilon_n} \int_{\partial E_i^{(n)}} \int_{|w| \leqslant 1} \int_{|y| \leqslant 1} \min\{|w_d|, |y_d|\} \, \mathrm{d}y \, \mathrm{d}w \, \mathrm{d}x_{1:d-1}$$

$$= \frac{|\partial E_i^{(n)}|_{\mathcal{H}^{d-1}} V}{2\epsilon_n}$$

$$= \frac{|\partial E_i|_{\mathcal{H}^{d-1}} V}{2\epsilon_n} + O(1).$$

Hence

$$\mathbb{E}\Xi_{ij}\Xi_{ik} = \frac{|\partial E|_{\mathcal{H}^{d-1}} V}{\epsilon_n} + O(1).$$

Collecting terms implies the result of the theorem. □

If one defines

$$\kappa_1 = 4|\partial E|_{\mathcal{H}^{d-1}} V$$

$$\kappa_2 = 2TV(\mu; \rho, \eta)$$

then we can conclude the asymptotic expansion given in § 2.3.

**Acknowledgements**

## Appendix A. Proof of Proposition 2.2

*Proof of Proposition 2.2.* For the first part assume the support of $\varphi$ is contained in $[0, M]$ and choose $N \leqslant M$ such that $\varphi(t) \geqslant \varphi(0)/2 > 0$ for all $0 \leqslant t \leqslant N$ and let $0 < \delta < N$. First consider $|x| \geqslant N - \delta$. If $|x - z| < \delta$ then

$$|z| \leqslant |z - x| + |x| \leqslant \delta + |x| = |x| \left(1 + \frac{\delta}{|x|}\right) \leqslant |x| \left(1 + \frac{\delta}{N - \delta}\right).$$

Set $c_\delta = 1$ and $\alpha_\delta = 1 + \delta/N - \delta$, then as $\varphi$ is decreasing, we have

$$\eta(z) = \varphi(|z|) \geqslant \varphi(\alpha_\delta|x|) = c_\delta \eta(\alpha_\delta x).$$

Now if $|x| \leqslant N - \delta$ then $|z| \leqslant N \leqslant M$ and, therefore,

$$\eta(z) = \phi(|z|) \geqslant \varphi(|x|) - L\,||x| - |z|| \geqslant \varphi(|x|) - L|x - z| \geqslant \varphi(|x|) - L\delta$$

$$= \varphi(|x|) \left(1 - \frac{L\delta}{\varphi(|x|)}\right)$$

where $L$ is the Lipschitz constant of $\varphi$. By definition of $N$ we have that $\varphi(|x|) \geqslant \varphi(0)/2$ and, therefore,

$$\eta(z) \geqslant \varphi(|x|)\left(1 - 2L\delta/\varphi(0)\right).$$

Hence $\eta(z) \geqslant c_\delta \eta(x)$ where $c_\delta = 1 - 2L\delta/\varphi(0)$ and $\alpha_\delta = 1$.

For the second case let $B(0, 2m) \subset E \subset B(0, M)$ and $\delta^* = \operatorname{dist}_H(\partial E, B(0, m))$ where $\operatorname{dist}_H$ is the Hausdorff distance. Clearly, $\delta^* \geqslant m > 0$. Let

$$\partial_{\delta^*} E = \left\{ x \in E \cup \partial E \; : \; \operatorname{dist}_H(x, \partial E) \leqslant \frac{\delta^*}{2} \right\}.$$

Note that if $x \in \partial_{\delta^*} E$ then $\operatorname{dist}_H(x, B(0, m)) \geqslant \delta^*/2$. For any $x \in \partial_{\delta^*} E$ there exists a unique (by convexity) $\beta_x \geqslant 1$ such that $\beta_x x \in \partial E$. Furthermore, $\beta_x = |\beta_x x|/|x| \leqslant M/m$.

Let $\delta \leqslant \delta^*$ and pick $z \in \partial_\delta^+ := \{ x \in E^c \; : \; \operatorname{dist}_H(z, \partial E) \leqslant \delta \}$ then for any $x \in B(z, \delta) \cup E$ we have

$$
\begin{aligned}
\beta_x - 1 &= \frac{1}{|x|} \left( \beta_x |x| - |z| + |z| - |x| \right) \\
&\leqslant \frac{1}{m} \left( |\beta_x - x| + |z - x| \right) \\
&\leqslant \frac{1}{m} \left( |\beta_x - x| + \delta \right).
\end{aligned}
$$

Now we construct the triangle given in figure 5. Applying the cosine formula one has,

$$
\begin{aligned}
\cos(\theta) &= \frac{|z|^2 + |x|^2 - |z - x|^2}{2|x||z|} \\
&= 1 + \frac{|z|^2 + |x|^2 - |z - x|^2 - 2|x||z|}{2|x||z|} \\
&\geqslant 1 - \frac{|z - x|^2}{2|x||z|} \\
&\geqslant 1 - \frac{\delta^2}{2m^2}.
\end{aligned}
$$

And

$$
\begin{aligned}
|\beta_x x - z|^2 &= |z|^2 + |\beta_x x|^2 - 2|z||\beta_x x| \cos(\theta) \\
&\leqslant |z|^2 + |\beta_x x|^2 - 2|z||\beta_x x| + \frac{|z||\beta_x x|\delta^2}{m^2} \\
&\leqslant \frac{|z||\beta_x x|\delta^2}{m^2} \\
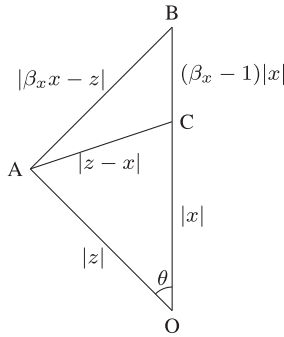&\leqslant \frac{M^3 \delta^2}{m^3}.
\end{aligned}
$$

Figure 5. Bound for $|\beta_x x - z|$ in the proof of Proposition 2.2.

Therefore,

$$\beta_x - 1 \leqslant \frac{\delta}{m}\left(\left(\frac{M}{m}\right)^{3/2} + 1\right).$$

Take $\alpha_\delta = \sup_{z \in \partial_\delta^+ E} \sup_{x \in B(z,\delta) \cap E} \beta_x$. Then $1 \leqslant \alpha_\delta \leqslant 1 + \delta/m((M/m)^{3/2} + 1) \to 1$ and by construction for any $x \in B(z,\delta) \cap E$, we have $\alpha x \notin E$. This implies $\eta(\alpha_\delta x) = 0$ and, therefore, $\eta(z) \geqslant \eta(\beta_x x)$.

The other cases, $x \in B(z,\delta) \cap E^c$ and when $z \in E$ or $\operatorname{dist}_H(z,\partial E) > \delta$ are trivial. $\qquad\square$

## Appendix B. Proof of Corollary 2.4

*Proof of Corollary* 2.4. The compactness property holds analogously to proposition 4.1. For the $\Gamma$-convergence, we let $\eta = \varphi \circ \pi = \varphi(|x|)$. Since $\varphi$ is in $L^1([0,\infty))$ (it is bounded, measurable and with compact support) then we can approximate $\varphi$ by $\varphi_k$ a monotonically increasing sequence (in $k$) of functions such that $0 \leqslant \varphi_k \leqslant \varphi$, $\varphi_k \to \varphi$ pointwise, $\varphi_k$ is Lipschitz, decreasing and $\varphi_k > 0$. By proposition 2.2 and theorem 2.3 for any $\mu \in L^1(X)$ there exists a sequence $\mu_n$ such that $\mu_n \to \mu$ in $TL^1$ and $\limsup_{n\to\infty} \mathcal{E}_n(\mu_n; \varphi_k \circ \pi) \leqslant \mathcal{E}_\infty(\mu; \varphi_k \circ \pi)$. Therefore,

$$\limsup_{n\to\infty} \mathcal{E}_n(\mu_n; \varphi \circ \pi)$$

$$\geqslant \limsup_{n\to\infty} \mathcal{E}_n(\mu_n; \varphi_k \circ \pi)$$

$$\geqslant \mathcal{E}_\infty(\mu; \varphi_k \circ \pi)$$

$$\to \mathcal{E}_\infty(\mu; \varphi \circ \pi) \quad \text{as } k \to \infty \text{ by the monotone convergence theorem.}$$

Similarly, for the liminf inequality, we take a sequence $\varphi_k$ monotonically decreasing sequence of functions such that $\varphi_k \geqslant \varphi$, $\varphi_k \to \varphi$ pointwise, $\varphi_k$ is Lipschitz,

decreasing and with compact support. Then for any $\mu_n \to \mu$ in $TL^1$ we have

$$\liminf_{n\to\infty} \mathcal{E}_n(\mu_n; \varphi \circ \pi)$$

$$\leqslant \liminf_{n\to\infty} \mathcal{E}_n(\mu_n; \varphi_k \circ \pi)$$

$$= \mathcal{E}_\infty(\mu; \varphi_k \circ \pi)$$

$$\to \mathcal{E}_\infty(\mu; \varphi \circ \pi) \quad \text{as } k \to \infty \text{ by the monotone convergence theorem.}$$

This completes the proof. □

## References

1   G. Alberti and G. Bellettini. A nonlocal anisotropic model for phase transitions: asymptotic behaviour of rescaled energies. *Eur. J. Appl. Math. Eur J Appl Math* **3** (1998), 527–560.
2   L. Ambrosio and A. Pratelli. Existence and stability results in the $L^1$ theory of optimal transportation. In *Optimal Transportation and Applications*, volume 1813 of *Lecture Notes in Mathematics*, L. A. Caffarelli and S. Salsa (eds), pp. 123–160 (Berlin, Heidelberg: Springer, 2003).
3   R. K. Ando and T. Zhang. Learning on graph with Laplacian regularization. *Advances in neural information processing systems* **19** (2007), 25–32.
4   H. Attouch, G. Buttazzo and G. Michaille. *Variational Analysis in Sobolev and BV Spaces: Applications to PDE's and Optimization*. MPS-SIAM Series on Optimization (Philadelphia: Society for Industrial and Applied Mathematics, 2006).
5   A. Baldi. Weighted BV functions. *Houst. J. Math.* **27** (2001), 683–705.
6   A. L. Bertozzi and A. Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale. Model. Simul.* **10** (2012), 1090–1118.
7   S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang. Complex networks: structure and dynamics. *Phys. Rep.* **424** (2006), 175–308.
8   A. Braides. Γ-*Convergence for Beginners* (New York: Oxford University Press, 2002).
9   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 25*, P. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger (eds), pp. 1385–1393 (Lake Tahoe: Curran Associates, Inc., 2012).
10  X. Bresson, T. Laurent, D. Uminsky and J. von Brecht. An adaptive total variation algorithm for computing the balanced cut of a graph. arXiv:1302.2717, (2013).
11  N. Bridle and X. Zhu. *p*-voltages: Laplacian regularization for semi-supervised learning on high-dimensional data. In *Eleventh Workshop on Mining and Learning with Graphs (MLG2013)* (2013).
12  A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener. Graph structure in the web. *Comput. Netw.* **33** (2000), 309–320.
13  T. Bühler and M. Hein. Spectral clustering based on the graph *p*-Laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 81–88, (2009).
14  G. Caldarelli. *Scale Free Networks: Complex Webs in Nature and Technology* (New York: Oxford University Press, 2007).
15  A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging. Vis.* **40** (2011), 120–145.
16  T. Champion, L. De Pascale and P. Juutinen. The ∞-Wasserstein distance: Local solutions and existence of optimal transport maps. *SIAM J. Math. Anal.* **40** (2008), 1–20.
17  G. Dal Maso. *An Introduction to Γ-Convergence* (New York: Springer, 1993).
18  L. Danon, A. Diaz-Guilera, J. Duch and A. Arenas. Comparing community structure identification. *J. Stat. Mech: Theory Exp.* **2005** (2005).
19  A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright and M. I. Jordan. Asymptotic behavior of $\ell_p$-based Laplacian regularization in semi-supervised learning. In *29th Annual Conference on Learning Theory*, pp. 879–906 (2016).

20    L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions* (Boca Raton: CRC Press, 1992).

21    M. Faloutsos, P. Faloutsos and C. Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM Computer Communication Review*, vol. 29, pp. 251–262 (1999).

22    S. Fortunato. Community detection in graphs. *Phys. Rep.* **3–5** (2010), 75–174.

23    W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Math.* **177** (1996), 113–161.

24    C. Garcia-Cardona, A. Flenner and A. G. Percus. Multiclass semi-supervised learning on graphs using Ginzburg–Landau functional minimization. In *Pattern Recognition Applications and Methods*, pp. 119–135 (Lisbon, Springer, 2015).

25    N. García Trillos and D. Slepčev. Continuum limit of Total Variation on point clouds. *Arch. Ration. Mech. Anal.* 1–49, 2015.

26    N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in $\infty$-transportation distance. *Can. J. Math.* **67** (2015), 1358–1383.

27    N. García Trillos, D. Slepčev, J. von Brecht, T. Laurent and X. Bresson. Consistency of cheeger and ratio graph cuts. arXiv:1411.6590 (2014).

28    J. F Hennawi and J. X Prochaska. Quasars probing quasars. II. The anisotropic clustering of optically thick absorbers around quasars. *Astrophys. J.* **655** (2007).

29    H. Hu, Y. van Gennip, B. Hunter, A. L. Bertozzi and M. A. Porter. Multislice modularity optimization in community detection and image segmentation. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, pp. 934–936, (2012).

30    H. Jylhä. The $L^\infty$ optimal transport: infinite cyclical monotonicity and the existence of optimal transport maps. *Calc. Var. Partial. Differ. Equ.* **52** (2015), 303–326.

31    B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* J. Langford and J. Pineau. (eds), pp. 513–520. (Edinburgh, Omnipress, 2012).

32    G. Leoni. *A First Course in Sobolev Spaces*, vol. 105 (Providence: American Mathematical Society, 2009).

33    X.-Z. Li, J.-F. Wang, W.-Z. Yang, Z.-J. Li and S.-J. Lai. A spatial scan statistic for nonisotropic two-level risk cluster. *Stat. Med.* **31** (2012), 177–187.

34    E. V. Linder, M. Oh, T. Okumura, C. G. Sabiu and Y.-S. Song. Cosmological constraints from the anisotropic clustering analysis using BOSS DR9. *Phys. Rev. D* **89** (2014), 063525

35    D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms* (Cambridge: Cambridge University Press, 2002).

36    V. Maz'ya. *Sobolev Spaces: With Applications to Elliptic Partial Differential Equations* (Berlin: Springer, 2010).

37    L. Modica. The gradient theory of phase transitions and the minimal interface criterion. *Arch. Ration. Mech. Anal.* **98** (1987), 123–142.

38    M. Penrose. *Random Geometric Graphs* (New York: Oxford University Press, 2003).

39    M. A. Porter, J.-P. Onnela and P. J. Mucha. Communities in networks. *Notices of the AMS* (2009).

40    S. H. Strogatz. Exploring complex networks. *Nature* **410**(6825) (2001), 268–276.

41    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** (2015), 2444–2474.

42    Y. Van Gennip and A. L. Bertozzi. Γ-convergence of graph Ginzburg–Landau functionals. *Adv. Differ. Equ.* **17**(11–12) (2012), 1115–1180.

43    Y. van Gennip, B. Hunter, R. Ahn, P. Elliott, K. Luh, M. Halvorson, S. Reid, M. Valasik, J. Wo, G. Tita, A. Bertozzi and P. Brantingham. Community detection using spectral clustering on sparse geosocial data. *SIAM J. Appl. Math.* **73** (2013), 67–83.

44    Y. van Gennip, N. Guillen, B. Osting and A. Bertozzi. Mean curvature, threshold dynamics, and phase field theory on finite graphs. *Milan J. Math.* **82** (2014), 3–65.

45    C. Villani. *Topics in Optimal Transportation*. Graduate Studies in Mathematics. American Mathematical Society, cop., (2003).

46    S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications* (Cambridge: Cambridge University Press, 1994).

47    K. J. Worsley, M. Andermann, T. Koulis, D. MacDonald and A. C. Evans. Detecting changes in nonisotropic images. *Hum. Brain. Mapp.* **8** (1999), 98–101.

48    X. Zhou and M. Belkin. Semi-supervised learning by higher order regularization. In *International Conference on Artificial Intelligence and Statistics*, pp. 892–900 (2011).

49    D. Zhou and B. Schölkopf. *Pattern Recognition: 27th DAGM Symposium, Vienna, Austria, 31 August–2 September 2005. Proceedings*, chapter Regularization on Discrete Spaces, pp. 361–368. (Berlin Heidelberg: Springer, 2005).

50    X. Zhou and N. Srebro. Error analysis of Laplacian eigenmaps for semi-supervised learning. In *AISTATS*, pp. 901–908, (2011).