

# Analysis of $p$ -Laplacian Regularization in Semi-Supervised Learning

Dejan Slepčev<sup>1</sup> and Matthew Thorpe<sup>2</sup>

<sup>1</sup>Department of Mathematical Sciences,  
Carnegie Mellon University,  
Pittsburgh, PA 15213, USA

<sup>2</sup>Department of Applied Mathematics and Theoretical Physics,  
University of Cambridge,  
Cambridge, CB3 0WA, UK

October 2017

## Abstract

We investigate a family of regression problems in a semi-supervised setting. The task is to assign real-valued labels to a set of  $n$  sample points, provided a small training subset of  $N$  labeled points. A goal of semi-supervised learning is to take advantage of the (geometric) structure provided by the large number of unlabeled data when assigning labels. We consider random geometric graphs, with connection radius  $\varepsilon(n)$ , to represent the geometry of the data set. Functionals which model the task reward the regularity of the estimator function and impose or reward the agreement with the training data. Here we consider the discrete  $p$ -Laplacian regularization.

We investigate asymptotic behavior when the number of unlabeled points increases, while the number of training points remains fixed. We uncover a delicate interplay between the regularizing nature of the functionals considered and the nonlocality inherent to the graph constructions. We rigorously obtain almost optimal ranges on the scaling of  $\varepsilon(n)$  for the asymptotic consistency to hold. We prove that the minimizers of the discrete functionals in random setting converge uniformly to the desired continuum limit. Furthermore we discover that for the standard model used there is a restrictive upper bound on how quickly  $\varepsilon(n)$  must converge to zero as  $n \rightarrow \infty$ . We introduce a new model which is as simple as the original model, but overcomes this restriction.

**Keywords and phrases.**  $p$ -Laplacian, regression, asymptotic consistency, asymptotics of discrete variational problems, Gamma-convergence, PDE on graphs, nonlocal variational problems

**Mathematics Subject Classification.** 49J55, 49J45, 62G20, 35J20, 65N12

## 1 Introduction

Due to its applicability across a large spectrum of problems semi-supervised learning (SSL) is an important tool in data analysis. It deals with situations when one has access to relatively few labeled

points but potentially a large number of unlabeled data. We assume that we are given  $N$  labeled points  $\{(x_i, y_i) : i = 1, \dots, N, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}$  and  $n - N$  points  $x_i, i = N + 1, \dots, n$  drawn from a fixed, but unknown measure,  $\mu$  supported in a compact subset of  $\mathbb{R}^d$ . The goal is to assign labels to the unlabeled points, while taking advantage of the information provided by the unlabeled points when designing the estimator. In particular the unlabeled points carry information on the structure of  $\mu$ , such as the geometry of its support, which can lead to better estimators. To access the information on  $\mu$  in a way that carries over to high dimensions, we build a graph whose vertices are data points and connect them if they are close enough, that is if they are within some distance  $\varepsilon > 0$ . More generally the edge weights are prescribed by using a decreasing function  $\eta : [0, \infty) \rightarrow [0, \infty)$  with  $\lim_{r \rightarrow \infty} \eta(r) = 0$ . For fixed scale  $\varepsilon > 0$  we set the weights to be

$$W_{ij} = \eta_\varepsilon(|x_i - x_j|)$$

where  $\eta_\varepsilon = \frac{1}{\varepsilon^d} \eta(\cdot/\varepsilon)$ .

The regression problem is to find an estimator  $u : \Omega_n := \{x_i : i = 1, \dots, n\} \rightarrow \mathbb{R}$  which agrees with preassigned labels. To solve the regression problem one considers objective functions which penalize the lack of smoothness of  $u$  and take the structure of the graph into account. In particular here we consider the functionals which generalize the graph Laplacian, namely the graph  $p$ -Laplacian. A particular objective function we consider is

$$(1) \quad \mathcal{E}_n^{(p)}(f) = \frac{1}{\varepsilon_n^p n^2} \sum_{i,j=1}^n W_{ij} |f(x_i) - f(x_j)|^p.$$

We consider minimizing  $\mathcal{E}_n^{(p)}(f)$  under the constraint that

$$(2) \quad f(x_i) = y_i \text{ for all } i = 1, \dots, N.$$

A numerically computed example of the minimizer of the functional is shown on Figure 1(a).

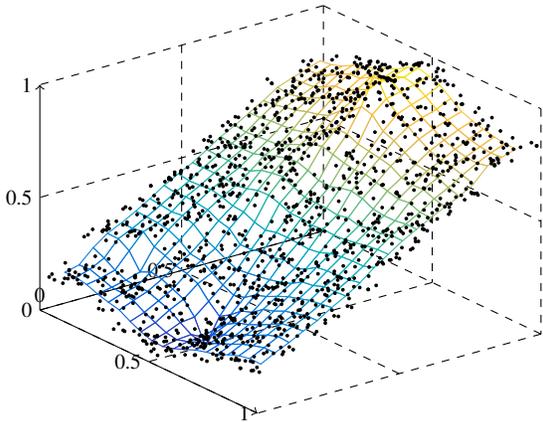
We investigate the asymptotic behavior in the limit when the number of unlabeled data goes to infinity, which is consistent with semi-supervised learning paradigm of having few labeled points and an abundance of unlabeled data. As  $n \rightarrow \infty$ ,  $\varepsilon(n) \rightarrow 0$  to increase the resolution and limit the computational cost. Namely as  $\varepsilon(n)$  is the length scale over which the information on  $\mu$  is averaged, taking  $\varepsilon(n)$  to zero insures that the finer scales of  $\mu$  are resolved as more data points become available.

We assume that  $\mu$  has density  $\rho$  which has a positive lower bound on an open set  $\Omega$  and is zero otherwise. While in this paper we consider data which are distributed in the set of full dimension, we remark that there are no essential difficulties to extend the results to manifold setting, namely one where  $\mu$  is a measure supported on compact manifold  $\mathcal{M}$  of dimension  $d$  embedded in  $\mathbb{R}^D$ . such extension has already been done for related problems concerning the graph laplacian [29], where the modification of background results (such as optimal transportation estimates) has been carried out.

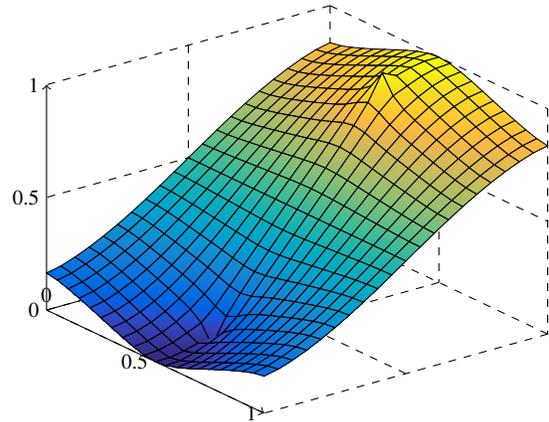
The continuum limiting problem corresponds to minimizing

$$(3) \quad \mathcal{E}_\infty^{(p)}(f) = \sigma \int_\Omega |\nabla f(x)|^p \rho^2(x) dx$$

where  $\sigma$  is a constant that depends on  $\eta$ , subject to constraint that  $f(x_i) = y_i$  for  $i = 1, \dots, N$ . A numerically computed minimizer of the functional is shown in Figure 1(b). Finiteness of  $\mathcal{E}_\infty^{(p)}(f)$  implies that  $f$  lies in the Sobolev space  $W^{1,p}(\Omega)$ . For the constraints to make sense it is needed that



(a) Minimizer of (1) under constraint (2) for  $\varepsilon = 0.058$  and  $\eta = \mathbb{1}_{[0,1]}$  and  $n = 1280$ . The grid is to aid visualization.



(b) Minimizer of the continuum functional (3) under constraint (2).

Figure 1: 2D numerical experiment for measure  $\mu$  with density one on  $[0, 1]^2$ , training data  $x_1 = (0.2, 0.5)$  and  $x_2 = (0.8, 0.5)$  and labels  $y_1 = 0$  and  $y_2 = 1$ , and  $p = 4$ .

pointwise evaluation of functions is well defined, which is the case only if  $p > d$ , when the Sobolev embedding ensures that functions in  $W^{1,p}$  are continuous. When  $p \leq d$  and  $d > 1$ , one cannot expect to be able to impose pointwise data. Indeed spikes were observed in discrete models with graph-Laplacian-based regularizations (that is for  $p = 2$ ) by Nadler, Srebro, and Zhou in [41] who also argued that they arise since there exist functions with arbitrarily small energy  $\mathcal{E}_n^{(p)}(f)$ , for  $p = 2$ , which agree with labels on the training set. In [16] El Alaoui, Cheng, Ramdas, Wainwright, and Jordan go a step further and suggest  $p = d$  as the transition point between the regime where spikes appear and where solutions are “smooth”. They argue, based on Sobolev embedding theorem, that for  $p \leq d$  the minimizers of  $\mathcal{E}_n^{(p)}(f)$  can develop spikes as  $n \rightarrow \infty$ , while for  $p > d$  they should not develop spikes (the authors consider  $p \geq d + 1$ , but the same argument applies for  $p > d$ ). The authors also argue that for data purposes taking  $p > d$  and close to  $d$  is optimal since as  $p \rightarrow \infty$  the solution forgets the information provided by the unlabeled points and only depends on the labeled ones.

Our initial goal was to verify the conclusions of [16]. More precisely to show that minimizers of  $\mathcal{E}_n^{(p)}(f)$ , constrained to agree with the provided labels, converge, in the appropriate topology, to minimizers of  $\mathcal{E}_\infty^{(p)}(f)$ , which also respect the labels, as  $n \rightarrow \infty$  when  $p > d$ , and that they develop spikes when  $p \leq d$ . However we discovered an additional phenomenon, namely that the undesirable spikes in the minimizers to graph  $p$ -Laplacian can occur even when  $p > d$ .

Namely [16] shows pointwise convergence of the form

$$\lim_{\varepsilon \rightarrow 0} \lim_{n \rightarrow \infty} \mathcal{E}_n^{(p)}(f) = \mathcal{E}_\infty^{(p)}(f),$$

when  $f$  is smooth enough. However considering a fixed function  $f$  is not sufficient to conclude that the constrained minimizers of  $\mathcal{E}_n^{(p)}$  converge to constrained minimizers of  $\mathcal{E}_\infty^{(p)}$ . In fact answering that question requires a set of tools from applied analysis which we discuss below. We show, roughly speaking, that for  $d \geq 3$  the convergence of minimizers holds if and only if

$$(4) \quad \left(\frac{1}{n}\right)^{\frac{1}{p}} \gg \varepsilon_n \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d}} \quad \text{as } n \rightarrow \infty.$$

The lower bound above is related to the connectivity of the graph constructed and was well understood, [32, 33]. Our lower bounds for  $d = 1, 2$  contain additional correction terms and are not optimal. Our upper bound implies that the models are in fact not consistent for a large family of scalings of  $\varepsilon$  on  $n$  that were thus far thought to ensure consistency (namely for  $1 \gg \varepsilon_n \gg n^{-1/p}$ ). Our work indicates that careful analytical approaches are needed and are in fact capable of providing precise information on asymptotic consistency of algorithms.

In the ‘‘ill-posed’’ regime  $\varepsilon_n^p n \rightarrow \infty$ , under the usual connectivity requirement (which when  $d \geq 3$  reads  $\varepsilon_n^d \frac{n}{\log n} \rightarrow \infty$ ), we are still able to establish the asymptotic behavior of algorithms. Namely we show that minimizers of  $\mathcal{E}_n^{(p)}(f)$  with constraints converge, along subsequences, as  $n \rightarrow \infty$  and  $\varepsilon_n \rightarrow 0$  to a minimizer of  $\mathcal{E}_\infty^{(p)}(f)$  without constraints. That is, the labels are forgotten in the limit as  $n \rightarrow \infty$ . This explains why, for large  $n$ , minimizers of  $\mathcal{E}_n^{(p)}$  are ‘spikey’. The need to consider subsequences in the limit is due to the fact that minimizers of  $\mathcal{E}_\infty^{(p)}(f)$  without constraints are nonunique.

While the degeneracy of the problem when  $p \leq d$  was known, [16], we believe that degeneracy when  $p > d$  and  $\varepsilon_n^p n \rightarrow \infty$  is a new and at first surprising result. The heuristic explanation for the appearance of spikes is that the discrete  $p$ -Laplacian does not share the regularizing properties of the continuum  $p$ -Laplacian. Namely the discrete  $p$ -Laplacian still involves averaging over the length scale  $\varepsilon$  and thus more closely resembles an integral operator (one in (14) to be precise). This allows high-frequency irregularities to form, without paying a high price in the energy. In particular, if we consider one labeled point taking the value 1, say  $f_n(x_1) = 1$ , while  $f_n(x_i) = 0$  for all  $i \geq 2$  then

$$\mathcal{E}_n^{(p)}(f_n) = \frac{2}{\varepsilon_n^p n^2} \sum_{j=2}^n \frac{1}{\varepsilon_n^d} \eta \left( \frac{|x_1 - x_j|}{\varepsilon_n} \right) = \frac{2}{\varepsilon_n^p n} \eta_{\varepsilon_n} * \mu_n(x_1) \rightarrow 0$$

as  $n \rightarrow \infty$ , when  $\varepsilon_n^p n \rightarrow \infty$ . Note that  $f_n$  exhibits degeneracy while  $\mathcal{E}_n^{(p)}(f_n) \rightarrow 0$ .

In addition to the constrained problem above we also consider the problem where the agreement with the labels provided is imposed through a penalty term. Our results and analysis are analogous.

Using the insights of our analysis, we define a new model which is quite similar to the original one, but for which the asymptotic consistency holds with only upper bound requirement being that  $\varepsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ .

To prove our results we use the tools of calculus of variations and optimal transportation. In particular we use the setup for convergence of objective functionals defined on graphs to their continuum limits developed in [32]. This includes the definition of the proper topology ( $TL^p$ ) to compare functionals defined on finite discrete objects (graphs) with their continuum limits. However the  $TL^p$  topology, which is an extension of the  $L^p$  topology, is not strong enough to ensure that the labels are preserved in the limit. For this reason we also need to consider a stronger topology, namely the one of uniform convergence. Proving the needed local regularity results for the discrete  $p$ -Laplacian (Lemma 4.1) and the compactness results needed to ensure the locally uniform convergence are the main technical contributions of the paper. We note that to the best of our knowledge, our results are the first where one proves (locally) uniform convergence of minimizers of nonlinear functionals in random discrete setting to the minimizers of the corresponding continuum functional.

We note that our results on asymptotic behavior of minimizers do not provide any error estimates for finite  $n$  and do not provide precise guidance on what  $\varepsilon$  would lead to best approximation. In Section 6, we numerically investigate prototypical examples in one and two dimensions to shed some light on

these issues. We numerically observe the predicted critical scalings for  $\varepsilon_n$  given in (4). We also numerically compare the results with our improved model (22). In investigating how precisely the observed error depends on  $\varepsilon$  we find that the error is smallest when  $\varepsilon$  is quite close to the connectivity radius on the graph. This is interesting and at first surprising. Rigorously explaining the phenomenon is in our opinion a valuable open problem.

The paper is organized as follows. We complete the introduction with a review on related works. In Section 2 we give a precise description of the problem with the assumptions and state the main results. Section 3 contains a brief overview of background results we use. This includes a description of the  $TL^p$  topology, which we use for discrete-to-continuum convergence, and a short overview of  $\Gamma$ -convergence and optimal transportation. Section 4 contains the proofs of the main results given in Section 2. In Section 5 we present an improved model that, while similar to the constrained problem for  $\mathcal{E}_n^{(p)}(f)$ , is asymptotically consistent with the desired limiting problem even when  $\varepsilon_n \rightarrow 0$  slowly as  $n \rightarrow \infty$ . We conclude the paper with 1D and 2D numerical experiments in Section 6.

## 1.1 Discussion of Related Works

The approach to semi-supervised learning using a weighted graph to represent the geometry of the unlabeled data and Laplacian based regularization was proposed by Zhu, Ghahramani, and Lafferty in [61]. It fits in the general theme of graph-Laplacian based approaches to machine learning tasks such as clustering, which are reviewed in [56]. See also [7] for a recent application to semi-supervised learning. Zhou and Schölkopf [59] generalized the regularizers of [61] to include a version of the graph  $p$ -Laplacian. The  $p$ -Laplacian regularization has also been used by Bühler and Hein in clustering problems [9], where values of  $p$  close to 1 are of particular interest due to connections with graph cuts. Graph based  $p$ -Laplacian regularization has found further applications in semi-supervised learning and image processing [17–19]. These papers also make the connection to the  $\infty$ -Laplacian, which is closely related to minimal Lipschitz extensions [13].

While the approach of [61] has found many applications it was pointed out by Nadler, Srebro and Zhou [41] that the estimator degenerates and becomes uninformative in  $d \geq 2$ , when the number of unlabeled data points  $n \rightarrow \infty$ . Almagir and von Luxburg [2] explored the  $p$ -resistances, the resulting distance on graphs, and connections to the  $p$ -Laplace regularization. Based on their analysis they suggested that  $p = d$  should be a good choice to prevent degeneracy in the  $n \rightarrow \infty$  limit. El Alaoui, Cheng, Ramdas, Wainwright, and Jordan [16] show that for  $p \leq d$  the problem degenerates as  $n \rightarrow \infty$  and spikes can occur. They argue that regularizations with high  $p \geq d + 1$  are sufficient to prevent the appearance of spikes as  $n \rightarrow \infty$ , and lead to a well-posed problem in the limit. Here we make part of their claims rigorous, namely that if  $p > d$  then the asymptotic consistency holds only if  $\varepsilon_n$  converges to zero sufficiently fast ( $\varepsilon_n n^p \rightarrow 0$  as  $n \rightarrow \infty$ ). If  $p > d$  and  $\varepsilon_n n^p \rightarrow \infty$  as  $n \rightarrow \infty$  we prove that the problem still degenerates as  $n \rightarrow \infty$  and that spikes occur. We also introduce a modification to the discrete problem (by modifying how the agreement with the assigned labels is imposed) which is well posed when  $p > d$  without the need for  $\varepsilon_n$  to converge to 0 quickly.

There are other ways to regularize the SSL regression problems which ensure that no spikes occur. Namely Belkin and Niyogi [4,5] consider estimators which are required to lie in the space spanned by a fixed number of eigenvectors of the graph Laplacian. Due to the smoothness of low eigenvectors of the Laplacian this prevents the formation of spikes. One can think of this approach in energy based setting where infinite penalty has been imposed on high frequencies. A softer, but still linear, way to do this is to consider (fractional) powers of the graph Laplacian, namely the regularity term  $J_n(u) = \langle cL_n^\alpha f, f \rangle$  where  $L_n$  is the graph Laplacian, and  $\alpha > 0$ . This regularization was studied by Belkin and Zhou [60]

who argue, again via regularity obtained by Sobolev embedding theorems, that taking  $\alpha > \frac{d}{2}$  prevents spikes. However Dunlop, Stuart, and the authors have discovered that a similar phenomenon to one described in this paper. Namely even when  $\alpha > \frac{d}{2}$  the limit may be degenerate, and spikes can occur, if  $\varepsilon_n$  converges to zero slowly, namely if  $\varepsilon_n^{2\alpha} n \rightarrow \infty$  as  $n \rightarrow \infty$ .

Our results fall in the class of asymptotic consistency results in machine learning. In general one is interested the asymptotic behavior of an objective function posed on a random sample of  $n$  points, and which also depends on a parameter  $\varepsilon$ ,  $E_{n,\varepsilon}(f_n)$  where  $f_n$  is a real valued function defined at sample points. The limit is considered as  $n \rightarrow \infty$  while  $\varepsilon_n \rightarrow 0$  at appropriate rate. The limiting problem is described by a continuum functional  $E_\infty(f)$  which acts on real valued functions supported on domains or manifolds. Also relevant is the (nonlocal) continuum problem,  $E_{\infty,\varepsilon}(f)$  which describes the limit  $n \rightarrow \infty$  while  $\varepsilon > 0$  is kept fixed.

The type of consistency that is needed for the conclusions, and the one we consider, is *variational consistency*, namely that minimizers of  $E_{n,\varepsilon_n}(f_n)$  converge to minimizers of  $E_\infty(f)$  as  $n \rightarrow \infty$  while  $\varepsilon_n \rightarrow 0$  at an appropriate rate. Proving such results includes choosing the right topology to compare the functions on discrete domain  $f_n$  with those on the continuum domain  $f$ .

Many works in the literature are interested in a simpler notion of convergence, namely that for a fixed, sufficiently smooth, continuum function  $f$  it holds that  $E_{n,\varepsilon_n}(f) \rightarrow E_\infty(f)$  as  $n \rightarrow \infty$  while  $\varepsilon_n \rightarrow 0$  at an appropriate rate, where by  $E_{n,\varepsilon_n}(f)$  we mean that the discrete functional is evaluated at the restriction of  $f$  to the data points. We call this notion of convergence *pointwise convergence*. A somewhat weaker notion of convergence is what we here call *iterated pointwise convergence*, namely considering  $\lim_{\varepsilon \rightarrow 0} \lim_{n \rightarrow \infty} E_{n,\varepsilon}(f)$ . Also relevant for the problems based on linear operators (namely the graph Laplacian) is *spectral convergence* which asks for the eigenvalues and eigenvectors of the discrete operator to converge to eigenvalues and eigenfunction of the continuum one. This notion of the convergence is typically sufficient for the kind of conclusions we are investigating (however our problems are nonlinear).

Pointwise (and similar notions of) convergence of graph Laplacians was studied by Belkin and Niyogi [6], Coifman and Lafon [12], Giné and Koltchinskii [35], Hein, Audibert and von Luxburg [37], Hein [36], Singer [47], and Ting, Huang, and Jordan [54]. Spectral convergence was studied in the works of Belkin and Niyogi [6] on the convergence of Laplacian eigenmaps, von Luxburg, Belkin, and Bousquet [57] and Pelletier and Pudlo [42] on graph Laplacians, and of Singer and Wu [48] on the connection graph Laplacian. In these works on spectral convergence either  $\varepsilon$  remains fixed as  $n \rightarrow \infty$  or  $\varepsilon(n) \rightarrow 0$  at an unspecified rate. The precise and almost optimal rates were obtained in [33] using variational methods. Further problems involve obtaining error estimates between discrete and continuum objects. Laplacians on discretized manifolds was studied by Burago, Ivanov and Kurylev [10] who obtain precise error estimates for eigenvalues and eigenvectors. Related results on approximating elliptic equations on point clouds have been obtained by Li and Shi [39], and Li, Shi, Sun, [40]. Error bounds for the spectral convergence of graph Laplacians have been considered by Wang [58] and García Trillos, Gerlach, Hein and one of the authors [29]. Regarding graph  $p$ -Laplacians, the authors of [16] obtain iterated pointwise convergence of graph  $p$ -Laplacians to the continuum  $p$ -Laplacian. Finally we mention that for a different type of problems, namely for nondominated sorting, of Calder, Esedoğlu, and Hero [11] have obtained uniform convergence of discrete solutions to the solution of a continuum Hamilton-Jacobi equation.

To obtain the results on variational convergence of  $\mathcal{E}_n^{(p)}$  to  $\mathcal{E}_\infty^{(p)}$  needed to fully explain the asymptotics of discrete regression problems we combine tools of calculus of variations (in particular  $\Gamma$ -convergence) and optimal transportation. This approach to asymptotics of problems posed on discrete

random samples was developed by García-Trillos and one of the authors [32,33]. In [32] they introduce the  $TLP$  topology for comparing the functions defined on the discrete sets to the ones defined in the continuum, and apply the approach to asymptotics of graph-cut based objective functions. We refer to this paper for a description of the rich background of the works that underpin the approach. In [33] the authors apply the approach to convergence of graph Laplacian based functionals. Consistency of  $k$ -means clustering for paths with regularization was recently studied by Theil, Johansen and Cade, and one of the authors [53], using a similar viewpoint. This technical setup has recently been used and extended to studies on modularity based clustering [15], Cheeger and ratio cuts [34], neighborhood graph constructions for graph cut based clustering [28], and classification problems [30, 52].

An alternative approach to related regression problems was developed by Fefferman and collaborators, Israel, Klartag and Luli, who look for a function of sufficient regularity, that extends a function  $f^\dagger : E \rightarrow \mathbb{R}$  to the whole of  $\mathbb{R}^d$  in such a way as to minimize the norm of the extension. They show that appropriate extensions exist and finding efficient constructions for  $f$ , for  $C^m$  regularity [21, 25, 26], and for Sobolev regularity [22–24]. In the context of machine learning this is a supervised learning problem and only makes use of the labeled data. In our context the problem is independent of  $\{x_i\}_{i=N+1}^n$  and does not use the geometry of the unlabeled data.

## 2 Setting and Main Results

Let  $\Omega$  be an open, bounded domain in  $\mathbb{R}^d$ . Let  $\{(x_i, y_i) : i = 1, \dots, N\}$  with  $x_i \in \Omega$  and  $y_i \in \mathbb{R}$  be a collection of distinct labeled points. Throughout the paper we consider  $N$  to be fixed. Considering a model where  $N$  grows is an interesting problem, which we do not address here. We consider  $\mu$  to be the measure representing the distribution of data. We assume that  $\text{supp}\mu = \bar{\Omega}$  and that  $\mu$  has density  $\rho$  with respect to Lebesgue measure. We assume that  $\rho$  is continuous and is bounded above and below by positive constants on  $\Omega$ .

We assume that unlabeled data,  $\{x_i\}_{i=N+1, \dots}$  are given by a sequence of iid samples of measure  $\mu$ . The empirical measure induced by data points is given by  $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ . Let  $G_n = (\Omega_n, E_n, W_n)$  be a graph with vertices  $\Omega_n = \{x_i : i = 1, \dots, n\}$ , edges  $E_n = \{e_{ij}\}_{i,j=1}^n$  and edge weights  $W_n = \{W_{ij}\}_{i,j=1}^n$ . For notational simplicity we will set  $W_{ij} = 0$  if there is no edge between  $x_i$  and  $x_j$ .

We assume the following structure on edge weights

$$(5) \quad W_{ij} = \eta_\varepsilon(|x_i - x_j|)$$

where  $\eta_\varepsilon(|x|) = \frac{1}{\varepsilon^d} \eta\left(\frac{|x|}{\varepsilon}\right)$ ,  $\eta : [0, \infty) \rightarrow [0, \infty)$  is a nonincreasing kernel and  $\varepsilon = \varepsilon_n$  is a scaling parameter depending on  $n$ . For example if  $\eta(|x|) = \mathbb{I}_{|x| \leq 1}$  then  $\eta_\varepsilon(|x|)$  is  $\frac{1}{\varepsilon^d}$  if  $|x| \leq \varepsilon$  and 0 otherwise. In this case vertices are only connected if they are closer than  $\varepsilon$ .

We consider two models: one where the agreement of the response with the training variables is imposed as a constraint and the other where it is imposed via a penalty. We call these models *constrained* and *penalized*.

In the constrained model we construct our estimator as the minimizer of

$$(6) \quad \mathcal{E}_n^{(p)}(f) = \frac{1}{\varepsilon_n^p} \frac{1}{n^2} \sum_{i,j=1}^n W_{ij} |f(x_i) - f(x_j)|^p$$

among  $\{f : \Omega_n \rightarrow \mathbb{R}\}$  which satisfy the constraint  $f(x_i) = y_i$  for all  $i = 1, \dots, N$ .

For technical reasons it is convenient to define the functional over all  $f$  and impose the constraint in the following way

$$(7) \quad \mathcal{E}_{n,con}^{(p)}(f) = \begin{cases} \frac{1}{\varepsilon_n^p} \frac{1}{n^2} \sum_{i,j=1}^n W_{ij} |f(x_i) - f(x_j)|^p & \text{if } f(x_i) = y_i \text{ for } i = 1, 2, \dots, N \\ \infty & \text{else.} \end{cases}$$

We now turn to the penalized formulation. For  $q > 0$  let

$$R^{(q)}(f) = \sum_{i=1}^N |y_i - f(x_i)|^q.$$

We define the estimator as the minimizer of

$$(8) \quad \mathcal{S}_n^{(p)}(f) = \mathcal{E}_n^{(p)}(f) + \lambda R^{(q)}(f)$$

where  $\lambda > 0$  is a tunable parameter.

We now introduce the continuum functionals that describe the limiting problems as  $n \rightarrow \infty$ . Let

$$(9) \quad \mathcal{E}_\infty^{(p)}(f) = \begin{cases} \sigma_\eta \int_\Omega |\nabla f(x)|^p \rho^2(x) dx & \text{if } f \in W^{1,p}(\Omega), \\ \infty & \text{else.} \end{cases}$$

For  $p > d$ , Sobolev functions  $f \in W^{1,p}$  are continuous and we can define

$$(10) \quad \mathcal{E}_{\infty,con}^{(p)}(f) = \begin{cases} \mathcal{E}_\infty^{(p)}(f) & \text{if } f \in W^{1,p}(\Omega) \text{ and } f(x_i) = y_i \text{ for } i = 1, \dots, N \\ \infty & \text{else.} \end{cases}$$

The constant  $\sigma_\eta$  above is defined, using  $e_1 = [1, 0, \dots, 0]^T$ , by

$$\sigma_\eta = \int_{\mathbb{R}^d} \eta(|x|) |x \cdot e_1|^p dx.$$

To describe the limit of the penalized model the large data limit we introduce

$$(11) \quad \mathcal{S}_\infty^{(p)}(f) = \mathcal{E}_\infty^{(p)}(f) + \lambda R^{(q)}(f).$$

We note that functionals (10) and (11) are lower semi-continuous with respect to the  $L^p$  norm. In addition, coercivity of both functionals follows from Sobolev embeddings. Coercivity and lower semi-continuity imply existence of minimizers, e.g. [27, Theorem 3.6]. Strict convexity implies that the minimizers are unique.

We are interested in asymptotic behavior of minimizers  $f_n$  of the discrete models, say  $\mathcal{E}_{n,con}^{(p)}$ . We say that  $\mathcal{E}_{n,con}^{(p)}$  is *asymptotically consistent* with  $\mathcal{E}_{\infty,con}^{(p)}$  if the minimizers  $f_n$  of  $\mathcal{E}_{n,con}^{(p)}$  converge as  $n \rightarrow \infty$  to a minimizer of  $\mathcal{E}_{\infty,con}^{(p)}$ . One should note topology of the convergence  $f_n \rightarrow f_\infty$  is not at this stage clear.

We observe that since  $f_n : \Omega_n \rightarrow \mathbb{R}$ , while  $f : \Omega \rightarrow \mathbb{R}$  this issue is nontrivial. We use the  $TL^p$  topology introduced in [32] precisely to compare functions defined on different domains in a topology consistent with  $L^p$  convergence. We define the convergence rigorously in Section 3.

Another issue is the rate at which  $\varepsilon_n$  is allowed to converge to zero. If  $\varepsilon_n \rightarrow 0$  too quickly then the graph becomes disconnected and hence it does not capture the geometry of  $\Omega$  properly. The connectivity threshold [43] is  $\varepsilon_n \sim \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$ . We require (when  $d \geq 3$ )  $\varepsilon_n \gg \left(\frac{\log n}{n}\right)^{\frac{1}{d}}$  which means that our lower bound is almost optimal. We discovered that if  $\varepsilon_n \rightarrow 0$  too slowly the discrete functional  $\mathcal{E}_{n,\text{con}}^{(p)}$  lacks sufficient regularity for the constraints to be preserved in the limit. The optimal upper bound on  $\varepsilon_n$  is discussed in Theorem 2.1.

We now state our assumptions needed for the main results.

- (A1)  $\Omega \subset \mathbb{R}^d$  is open, connected, bounded and with Lipschitz boundary;
- (A2) The probability measure  $\mu \in \mathcal{P}(\Omega)$  has continuous density  $\rho$  which is bounded above and below by strictly positive constants in  $\Omega$ ;
- (A3) There exists  $N$  labeled points:  $(x_i, y_i) \in \Omega \times \mathbb{R}$  for  $i = 1, \dots, N$ ;
- (A4) For  $i > N$  the data points  $x_i$ , are iid samples of  $\mu$ ;
- (A5) Let  $\varepsilon_n$  be a sequence converging to 0 satisfying the lower bound

$$\varepsilon_n \gg \begin{cases} \sqrt{\frac{\log \log n}{n}} & \text{if } d = 1 \\ \frac{(\log n)^{\frac{3}{4}}}{\sqrt{n}} & \text{if } d = 2 \\ \left(\frac{\log n}{n}\right)^{\frac{1}{d}} & \text{if } d \geq 3; \end{cases}$$

- (A6) The kernel profile  $\eta : [0, \infty) \rightarrow [0, \infty)$  is non-increasing;
- (A7)  $\eta$  is positive and continuous at  $x = 0$ ;
- (A8) The integral  $\int_0^\infty \eta(t)|t|^{p+d} dt$  is finite (equivalently  $\sigma_\eta = \int_{\mathbb{R}^d} \eta(|w|)|w \cdot e_1|^p dw < \infty$ ).

The first main result of the paper is the following theorem. Its proof is presented in Section 4.

**Theorem 2.1** (Consistency of the constrained model). *Let  $p > 1$ . Assume  $\Omega$ ,  $\mu$ ,  $\eta$ , and  $x_i$  satisfy the assumptions (A1) - (A8). Let graph weights  $W_{ij}$  be given by (5). Let  $f_n$  be a sequence of minimizers of  $\mathcal{E}_{n,\text{con}}^{(p)}$  defined in (10). Then, almost surely, the sequence  $(\mu_n, f_n)$  is precompact in the  $TL^p$  metric. The  $TL^p$  limit of any convergent subsequence,  $(\mu_{n_m}, f_{n_m})$ , is of the form  $(\mu, f)$  where  $f \in W^{1,p}(\Omega)$ . Furthermore,*

(i) if  $n\varepsilon_n^p \rightarrow 0$  as  $n \rightarrow \infty$  then  $f$  is continuous and

(a)  $f_{n_m}$  converges locally uniformly to  $f$ , meaning that for any  $\Omega' \subset\subset \Omega$

$$\lim_{m \rightarrow \infty} \max_{\{k \leq n_m : x_k \in \Omega'\}} |f(x_k) - f_{n_m}(x_k)| = 0,$$

(b)  $f$  is a minimizer of  $\mathcal{E}_{\infty,\text{con}}^{(p)}$  defined in (10),

(c) the whole sequence  $f_n$  converges to  $f$  both in  $TL^p$  and locally uniformly;

(ii) if  $n\varepsilon_n^p \rightarrow \infty$  as  $n \rightarrow \infty$  then  $f$  is a minimizer of  $\mathcal{E}_{\infty}^{(p)}$  defined in (9).

We note that in case (i) assumption **(A5)** and  $n\varepsilon_n^p \rightarrow 0$  as  $n \rightarrow \infty$  imply that  $n^{-1/p} \gg \varepsilon \gg n^{-1/d}$  which is only possible if  $p > d$ . Therefore in case (i) we always have that functions  $f$  for which  $\mathcal{E}_\infty^{(p)}$  is finite are always continuous and thus it is possible to impose pointwise values of  $f$ , as needed to define  $\mathcal{E}_{\infty,con}^{(p)}$  in (10).

The result (i) establishes the asymptotic consistency of the discrete constrained model with the constrained continuum weighted  $p$ -Laplacian model.

While the result (ii) looks similar its interpretation is different. It shows that the model “forgets” the constraints in the limit. Namely  $\mathcal{E}_\infty^{(p)}$  only has the gradient term and no constraints! In particular its minimizers are constants over  $\Omega$ . What is happening is that  $f_n$  develops narrow spikes near the labeled points  $x_i$  and becomes nearly constant everywhere else. In the  $TL^p$  limit the spikes disappear.

This motivates referring to the scaling when  $n^p\varepsilon \rightarrow \infty$  as  $n \rightarrow \infty$  as the **degenerate** regime. On the other hand, we refer to the scaling of case (i) as the **well-posed** regime.

The other main result is the convergence in the penalized model. The proof is a straightforward extension of Theorem 2.1 in the special case  $N = 0$  (so that the constraint is not present). We include the proof in Section 4.2.

**Proposition 2.2.** *Let  $p > 1$ . Assume  $\Omega$ ,  $\mu$ ,  $\eta$ , and  $x_i$  satisfy the assumptions **(A1)**-**(A8)**. Let graph weights  $W_{ij}$  be given by (5). Let  $f_n$  be a sequence of minimizers of  $\mathcal{S}_n^{(p)}$  defined in (8). Then, almost surely, the sequence  $(\mu_n, f_n)$  is precompact in the  $TL^p$  metric. The  $TL^p$  limit of any convergent subsequence,  $(\mu_{n_m}, f_{n_m})$ , is of the form  $(\mu, f)$  where  $f \in W^{1,p}(\Omega)$ . Furthermore,*

(i) if  $n\varepsilon_n^p \rightarrow 0$  as  $n \rightarrow \infty$  then  $f$  is continuous and

(a)  $f_n$  converges locally uniformly to  $f$ , meaning that for any  $\Omega' \subset\subset \Omega$

$$\lim_{n \rightarrow \infty} \max_{\{k \leq n : x_k \in \Omega'\}} |f(x_k) - f_n(x_k)| = 0,$$

(b)  $f$  is a minimizer of  $\mathcal{S}_\infty^{(p)}$  defined in (10),

(c) the whole sequence  $f_n$  converges to  $f$  both in  $TL^p$  and locally uniformly;

(ii) if  $n\varepsilon_n^p \rightarrow \infty$  as  $n \rightarrow \infty$  then  $f$  is a minimizer of  $\mathcal{E}_\infty^{(p)}$  defined in (9).

Again the result of (i) is a consistency result, while (ii) shows that the penalization of the labels is lost in the limit.

*Remark 2.3.* The above results (Theorem 2.1 and Proposition 2.2) could also be extended to  $p = 1$ , in which case the limiting functional  $\mathcal{E}_\infty^{(1)}$  would be a weighted  $TV$  semi-norm  $\mathcal{E}_\infty^{(1)} = \sigma_\eta TV(\cdot; \rho)$  where

$$TV(f; \rho) = \sup \left\{ \int_\Omega f \operatorname{div} \phi \, dx : |\phi(x)| \leq \rho^2(x) \forall x \in \Omega, \phi \in C_c^\infty(\Omega; \mathbb{R}^d) \right\}.$$

A modification of the proofs contained here would prove the result, see also [32].

We recall that in Section 5 we propose an improved model that is well-posed when  $p > d$  without requiring that  $n\varepsilon^p \rightarrow 0$ .

### 3 Background Material

In an effort to make this paper more self-contained we briefly recall three key notions our work relies on. The first is  $\Gamma$ -convergence which is a notion of convergence of functionals developed for the analysis of sequences of variational problems. The second is the notion of optimal transportation, and the third is the  $TL^p$  space which we use to define the convergence of discrete functions to continuum functions.

#### 3.1 $\Gamma$ -Convergence

$\Gamma$ -convergence was introduced by De Giorgi in 1970's to study limits of variational problems. We refer to [8, 14] for an in depth introduction to  $\Gamma$ -convergence. Our application of  $\Gamma$ -convergence will be in a random setting.

**Definition 3.1** ( $\Gamma$ -convergence). *Let  $(Z, d)$  be a metric space and  $(\mathcal{X}, \mathbb{P})$  be a probability space. For each  $\omega \in \mathcal{X}$  the functional  $E_n^{(\omega)} : Z \rightarrow \mathbb{R} \cup \{\pm\infty\}$  is a random variable. We say  $E_n^{(\omega)}$   $\Gamma$ -converge almost surely on the domain  $Z$  to  $E_\infty : Z \rightarrow \mathbb{R} \cup \{\pm\infty\}$  with respect to  $d$ , and write  $E_\infty = \Gamma\text{-}\lim_{n \rightarrow \infty} E_n^{(\omega)}$ , if there exists a set  $\mathcal{X}' \subset \mathcal{X}$  with  $\mathbb{P}(\mathcal{X}') = 1$ , such that for all  $\omega \in \mathcal{X}'$  and all  $f \in Z$ :*

(i) (liminf inequality) for every sequence  $\{f_n\}_{n=1}^\infty$  converging to  $f$

$$E_\infty(f) \leq \liminf_{n \rightarrow \infty} E_n^{(\omega)}(f_n), \text{ and}$$

(ii) (recovery sequence) there exists a sequence  $\{f_n\}_{n=1}^\infty$  converging to  $f$  such that

$$E_\infty(f) \geq \limsup_{n \rightarrow \infty} E_n^{(\omega)}(f_n).$$

For ease of notation we will suppress the dependence of  $\omega$  on our functionals, that is we apply the above definition to  $E_n = \mathcal{E}_n^{(p)}$ . The almost sure statement in the above definition does not play a significant role in the proofs. Basically it is enough to consider the set of realizations of  $\{x_i\}_{i=1}^\infty$  such that the empirical measure converges weak\*. More precisely, we consider the set of realizations of  $\{x_i\}_{i=1}^\infty$  such that the conclusions of Theorem 3.3 hold.

The fundamental result concerning  $\Gamma$ -convergence is the following convergence of minimizers result. The proof can be found in [8, Theorem 1.21] or [14, Theorem 7.23].

**Theorem 3.2** (Convergence of Minimizers). *Let  $(Z, d)$  be a metric space and  $E_n : Z \rightarrow [0, \infty]$  be a sequence of functionals. Let  $f_n$  be a minimizing sequence for  $E_n$ . If the set  $\{f_n\}_{n=1}^\infty$  is precompact and  $E_\infty = \Gamma\text{-}\lim_n E_n$  where  $E_\infty : Z \rightarrow [0, \infty]$  is not identically  $+\infty$  then*

$$\min_Z E_\infty = \lim_{n \rightarrow \infty} \min_Z E_n.$$

Furthermore any cluster point of  $\{f_n\}_{n=1}^\infty$  is a minimizer of  $E_\infty$ .

The theorem is also true if we replace minimizers with almost minimizers.

We note that  $\Gamma$ -convergence is defined for functionals on a common metric space. The next section overviews the metric space we use to analyze the asymptotics of our semi-supervised learning models, in particular it allows us to go from discrete to continuum.

### 3.2 Optimal Transportation and Approximation of Measures

Here we recall the notion of optimal transportation between measures and the metric it introduces. Comprehensive treatment of the topic can be found in books of Villani [55] and Santambrogio [45].

Given  $\Omega$  is open and bounded, and probability measures  $\mu$  and  $\nu$  in  $\mathcal{P}(\overline{\Omega})$  we define the set  $\Pi(\mu, \nu)$  of transportation maps, or couplings, between them to be the set of probability measures on the product space  $\pi \in \mathcal{P}(\overline{\Omega} \times \overline{\Omega})$  whose first marginal is  $\mu$  and second marginal is  $\nu$ . We then define the  $p$ -optimal transportation distance (a.k.a.  $p$ -Wasserstein distance) by

$$d_p(\mu, \nu) = \begin{cases} \inf_{\pi \in \Pi(\mu, \nu)} \left( \int_{\Omega \times \Omega} |x - y|^p d\pi(x, y) \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty \\ \inf_{\pi \in \Pi(\mu, \nu)} \pi\text{-ess sup}_{(x, y)} |x - y| & \text{if } p = \infty. \end{cases}$$

If  $\mu$  has a density with respect to Lebesgue measure on  $\Omega$ , then the distance can be rewritten using transportation maps,  $T : \Omega \rightarrow \Omega$ , instead of transportation plans,

$$d_p(\mu, \nu) = \begin{cases} \inf_{\pi \in \Pi(\mu, \nu)} \left( \int_{\Omega} |x - T(x)|^p d\mu(x) \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty \\ \inf_{T: \mu = \nu} \mu\text{-ess sup}_x |x - T(x)| & \text{if } p = \infty. \end{cases}$$

where  $T_\# \mu = \nu$  means that the push forward of measure  $\mu$  by  $T$  is measure  $\nu$ , namely that  $T$  is Borel measurable and such that for all  $U \subset \overline{\Omega}$ , open,  $\mu(T^{-1}(U)) = \nu(U)$ .

When  $p < \infty$  the metric  $d_p$  metrizes the weak convergence of measures.

Optimal transportation plays an important role in comparing the discrete and continuum objects we study. In particular we use sharp estimates on the  $\infty$ -optimal transportation distance between a measure and the empirical measure of its sample. In the form below, for  $d \geq 2$ , they were established in [31], which extended the related results in [1, 38, 46, 49]. For  $d = 1$  the estimates are simpler, and follow from the law of iterated logarithms.

**Theorem 3.3.** *Let  $\Omega \subset \mathbb{R}^d$  be open, connected and bounded with Lipschitz boundary. Let  $\mu$  be a probability measure on  $\Omega$  with density (with respect to Lebesgue)  $\rho$  which is bounded above and below by positive constants. Let  $x_1, x_2, \dots$  be a sequence of independent random variables with distribution  $\mu$  and let  $\mu_n$  be the empirical measure. Then there exists a constants  $C \geq c > 0$  such that almost surely there exists a sequence of transportation maps  $\{T_n\}_{n=1}^\infty$  from  $\mu$  to  $\mu_n$  such that*

$$c \leq \liminf_{n \rightarrow \infty} \frac{\|T_n - \text{Id}\|_{L^\infty(\Omega)}}{\delta_n} \leq \limsup_{n \rightarrow \infty} \frac{\|T_n - \text{Id}\|_{L^\infty(\Omega)}}{\delta_n} \leq C$$

where

$$\delta_n = \begin{cases} \sqrt{\frac{\log \log(n)}{n}} & \text{if } d = 1 \\ \frac{(\log n)^{\frac{3}{4}}}{\sqrt{n}} & \text{if } d = 2 \\ \frac{(\log n)^{\frac{1}{d}}}{n^{\frac{1}{d}}} & \text{if } d \geq 3. \end{cases}$$

### 3.3 The $TL^p$ Space

The discrete functionals we consider (e.g.  $\mathcal{E}_n^{(p)}$ ) are defined for functions  $f_n : \Omega_n \rightarrow \mathbb{R}$  where  $\Omega_n = \{x_i : i = 1, \dots, n\}$ , while the limit functional  $\mathcal{E}_\infty^{(p)}$  acts on functions  $f : \Omega \rightarrow \mathbb{R}$ , where  $\Omega$  is an open set. We can view  $f_n$  as elements of  $L^p(\mu_n)$  where  $\mu_n$  is the empirical measure of the sample  $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ . Likewise  $f \in L^p(\mu)$  where  $\mu$  is the measure with density  $\rho$  out of which the points are sampled from. One would like how to compare  $f$  and  $f_n$  in a way that is consistent with  $L^p$  topology. To do so we use the  $TL^p$  space was introduced in [32], where it was used to study the continuum limit of the graph total variation (that is  $\mathcal{E}_n^{(1)}$ ). Subsequent development of the  $TL^p$  space has been carried out in [33, 50, 51].

To compare the functions  $f_n$  and  $f$  above we need to take into account their domains, or more precisely to account for  $\mu$  and  $\mu_n$ . For that purpose the space of configurations is defined to be

$$TL^p(\Omega) = \{(\mu, f) : \mu \in \mathcal{P}(\overline{\Omega}), f \in L^p(\mu)\}.$$

The metric on the space is

$$d_{TL^p}^p((\mu, f), (\nu, g)) = \inf \left\{ \int_{\Omega \times \Omega} |x - y|^p + |f(x) - g(y)|^p d\pi(x, y) : \pi \in \Pi(\mu, \nu) \right\}$$

where  $\Pi(\mu, \nu)$  the set of transportation plans defined in Section 3.2. We note that the minimizing  $\pi$  exists and that  $TL^p$  space is a metric space, [32].

When  $\mu$  has a density with respect to Lebesgue measure on  $\Omega$ , then the distance can be rewritten using transportation maps,  $T$  instead of transportation plans,

$$d_{TL^p}^p((\mu, f), (\nu, g)) = \inf \left\{ \int_{\Omega} |x - T(x)|^p + |f(x) - g(T(x))|^p d\mu(x) : T_{\#}\mu = \nu \right\}.$$

This formula provides a clear interpretation of the distance in our setting. Namely to compare functions  $f_n : \Omega_n \rightarrow \mathbb{R}$  we define a mapping  $T_n : \Omega \rightarrow \Omega_n$  and compare the functions  $\tilde{f}_n = f_n \circ T_n$  and  $f$  in  $L^p(\mu)$ , while also accounting for the transport, namely the  $|x - T_n(x)|^p$  term.

We remark that  $TL^p(\overline{\Omega})$  space is not complete and that its completion was discussed in [32]. In the setting of this paper, since the corresponding measure is clear from context, we often say that  $f_n$  converges in  $TL^p$  to  $f$  as a short way to say that  $(\mu_n, f_n)$  converges in  $TL^p$  to  $(\mu, f)$ .

## 4 Regularity and Asymptotics of Discrete and Nonlocal Functionals

Here we present some of the key properties of the functionals involved that allow us to show the asymptotic consistency of Theorem 2.1. A fundamental new issue (compared to say [33]) is that constraints in  $\mathcal{E}_\infty^{(p)}$  are imposed pointwise on a set of  $\mu$  measure zero. [The reason that these constraints make sense is that for  $p > d$  the finiteness of  $\mathcal{E}_\infty^{(p)}(f)$  implies that  $f$  is continuous.] We note that the  $TL^p$  convergence used in [33] is not sufficient to imply that constraints are preserved. One needs a stronger convergence, like the uniform one. This raises the question on how to obtain the needed compactness of sequences  $f_n$ , that is how to show that uniform boundedness of  $\mathcal{E}_{n,con}^{(p)}(f_n)$  implies the existence of a (locally) uniformly converging subsequence. Our approach combines discrete and continuum regularity results. Namely we obtain in Lemma 4.1 a local control of oscillation of  $f_n$  over distances of order  $\varepsilon_n$ . In Lemma 4.2 we show that discrete functionals  $\mathcal{E}_{n,con}^{(p)}(f_n)$  control the values

of the associated nonlocal continuum functionals  $\mathcal{E}_{\varepsilon_n}^{(NL,p)}(\tilde{f}_n)$  (defined in (14) below) applied to an appropriate extrapolation  $\tilde{f}_n$  of  $f_n$ . A simple but important point is that the discrete functionals at fixed  $n$  are always closer to a nonlocal functional with nonlocality at scale  $\varepsilon_n$ , than to the limiting functional. The issue is that these nonlocal functionals do not share the regularizing properties of the limiting functional. However we show in Lemma 4.3 that control of the nonlocal energy is sufficient to provide regularity at scales larger than  $\varepsilon$ . Combining these estimates is enough to imply the compactness with respect to (locally) uniform convergence, Lemma 4.5.

**Lemma 4.1** (discrete regularity). *Let  $p > 1$ . Assume  $\Omega$ ,  $\mu$ ,  $\eta$ , and  $x_i$  satisfy the assumptions (A1) - (A8). Let graph weights  $W_{ij}$  be given by (5). Let  $\Omega_n = \{x_i\}_{i=1}^n$ . For any  $f_n : \Omega_n \rightarrow \mathbb{R}$ , we define  $\text{osc}_{\varepsilon}^{(n)}(f_n) : \Omega_n \rightarrow \mathbb{R}$  by*

$$\text{osc}_{\varepsilon}^{(n)}(f_n)(x_i) = \max_{z \in B(x_i, \varepsilon) \cap \Omega_n} f_n(z) - \min_{z \in B(x_i, \varepsilon) \cap \Omega_n} f_n(z).$$

*For any  $\alpha_0 > 0$ , with probability one, there exist  $n_0 > 0$  and  $C > 0$  (independent of  $n$ ) such that for any  $\alpha \geq \alpha_0$ , all  $n \geq n_0$ , all  $f_n : \Omega_n \rightarrow \mathbb{R}$ , and all  $k \in \{1, 2, \dots, n\}$*

$$\left( \text{osc}_{\alpha \varepsilon_n}^{(n)}(f_n)(x_k) \right)^p \leq C \alpha^p n \varepsilon_n^p \mathcal{E}_n^{(p)}(f_n),$$

*where  $\mathcal{E}_n^{(p)}$  is defined by (6).*

*Proof.* Let  $\tilde{\eta}(t) = a$  if  $0 \leq t < b$  and  $\tilde{\eta}(t) = 0$  where  $a$  and  $b$  are chosen such that  $\tilde{\eta} \leq \eta$ . We can furthermore choose  $b$  so that  $b \leq \alpha_0$ . For all  $k \in \{1, \dots, n\}$  let

$$\begin{aligned} \bar{f}_n(x_k) &= \max_{z \in B(x_k, \frac{b\varepsilon_n}{2}) \cap \Omega_n} f_n(z), & \bar{x}_k &\in \operatorname{argmax}_{z \in B(x_k, \frac{b\varepsilon_n}{2}) \cap \Omega_n} f_n(z), \\ \underline{f}_n(x_k) &= \min_{z \in B(x_k, \frac{b\varepsilon_n}{2}) \cap \Omega_n} f_n(z), & \underline{x}_k &\in \operatorname{argmin}_{z \in B(x_k, \frac{b\varepsilon_n}{2}) \cap \Omega_n} f_n(z). \end{aligned}$$

Note that  $\text{osc}_{\frac{b\varepsilon_n}{2}}^{(n)}(f_n)(x_k) = \bar{f}_n(x_k) - \underline{f}_n(x_k)$  and for all  $x \in B(x_k, \frac{b\varepsilon_n}{2}) \cap \Omega_n$

$$\begin{aligned} \text{(i)} \quad & \bar{f}_n(x_k) - f_n(x) \geq \frac{1}{2} \text{osc}_{\frac{b\varepsilon_n}{2}}(f_n)(x_k), \\ \text{or (ii)} \quad & f_n(x) - \underline{f}_n(x_k) \geq \frac{1}{2} \text{osc}_{\frac{b\varepsilon_n}{2}}(f_n)(x_k). \end{aligned}$$

Without a loss of generality we assume that (i) holds for at least half the points in  $B(x_k, \frac{b\varepsilon_n}{2}) \cap \Omega_n$ .

Then,

$$\begin{aligned}
\mathcal{E}_n^{(p)}(f_n) &\geq \frac{1}{\varepsilon_n^{p+d} n^2} \sum_{i,j=1}^n \tilde{\eta} \left( \frac{|x_i - x_j|}{\varepsilon_n} \right) |f_n(x_i) - f_n(x_j)|^p \\
&\geq \frac{1}{\varepsilon_n^{p+d} n^2} \sum_{j: |x_j - \bar{x}_k| \leq b\varepsilon_n} \tilde{\eta} \left( \frac{|\bar{x}_k - x_j|}{\varepsilon_n} \right) |f_n(x_j) - f_n(\bar{x}_k)|^p \\
(12) \quad &\geq \frac{a}{\varepsilon_n^{p+d} n^2} \sum_{j: |x_j - x_k| \leq \frac{b\varepsilon_n}{2}} |f_n(x_j) - f_n(\bar{x}_k)|^p, \quad \text{since } |x_k - \bar{x}_k| \leq \frac{b\varepsilon_n}{2} \\
&\geq \frac{a}{2^{p+1} \varepsilon_n^{p+d} n^2} \left( \text{osc}_{\frac{b\varepsilon_n}{2}}(f_n)(x_k) \right)^p \# \left\{ j : |x_j - x_k| \leq \frac{b\varepsilon_n}{2} \right\} \\
&= \frac{a}{2^{p+1} \varepsilon_n^{p+d} n} \left( \text{osc}_{\frac{b\varepsilon_n}{2}}(f_n)(x_k) \right)^p \mu_n \left( B \left( x_k, \frac{b\varepsilon_n}{2} \right) \right).
\end{aligned}$$

where  $\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ . Now, for a transport map  $T_n : \Omega \rightarrow \Omega_n$  from  $\mu$  to  $\mu_n$ , satisfying the conclusions of Theorem 3.3, we have

$$\begin{aligned}
(13) \quad \frac{1}{\varepsilon_n^d} \mu_n \left( B \left( x_k, \frac{b\varepsilon_n}{2} \right) \right) &= \frac{1}{\varepsilon_n^d} \int_{\Omega} \mathbb{I}_{\{|T_n(x) - x_k| \leq \frac{b\varepsilon_n}{2}\}} \rho(x) \, dx \\
&\geq \frac{\inf_{x \in \Omega} \rho}{\varepsilon_n^d} \int_{\Omega} \mathbb{I}_{\{|x - x_k| \leq \frac{b\varepsilon_n}{2} - \|T_n - \text{Id}\|_{L^\infty}\}} \rho(x) \, dx \\
&= \left( \inf_{x \in \Omega} \rho(x) \right) \text{Vol} \left( B \left( 0, \frac{b}{2} - \frac{\|T_n - \text{Id}\|_{L^\infty}}{\varepsilon_n} \right) \right).
\end{aligned}$$

We choose  $n_0$  such that for  $n \geq n_0$  it holds that  $\frac{\|T_n - \text{Id}\|_{L^\infty}}{\varepsilon_n} \leq \frac{b}{4}$ . Combining (12) and (13) gives

$$\left( \text{osc}_{\frac{b\varepsilon_n}{2}}(f_n)(x_k) \right)^p \leq \frac{2^{p+1} \varepsilon_n^p n \mathcal{E}_n^{(p)}(f_n)}{a \left( \inf_{x \in \Omega} \rho(x) \right) \text{Vol} \left( B \left( 0, \frac{b}{4} \right) \right)} =: C_1 \varepsilon_n^p n \mathcal{E}_n^{(p)}(f_n).$$

For  $\alpha > \alpha_0$ , using  $\alpha_0 \geq b$  and applying the triangle inequality  $\lfloor \frac{2\alpha}{b} \rfloor$  times, we obtain

$$\left( \text{osc}_{\alpha \varepsilon_n}(f_n)(x_k) \right)^p \leq C_1 \left( \left\lfloor \frac{2\alpha}{b} \right\rfloor + 1 \right)^p \varepsilon_n^p n \mathcal{E}_n^{(p)}(f_n) \leq C_1 \left( \frac{3\alpha}{b} \right)^p \varepsilon_n^p n \mathcal{E}_n^{(p)}(f_n)$$

which completes the proof.  $\square$

**Lemma 4.2** (discrete to nonlocal control). *Let  $p \geq 1$ . Assume  $\Omega$ ,  $\mu$ ,  $\eta$ , and  $x_i$  satisfy (A1) - (A8). Let graph weights  $W_{ij}$  be given by (5). Let constants  $a, b > 0$  be such that for  $\tilde{\eta}(|x|) = a$  for  $|x| \leq b$  and  $\tilde{\eta}(|x|) = 0$  otherwise it holds that  $\tilde{\eta} \leq \eta$ . Let  $T_n$  be a transport map satisfying the results of Theorem 3.3 and let  $\tilde{\varepsilon}_n = \varepsilon_n - \frac{2\|T_n - \text{Id}\|_{L^\infty}}{b}$ . Then there exists constants  $n_0 > 0$  and  $C > 0$  (independent of  $n$  and  $f_n$ ) such that for all  $n \geq n_0$*

$$\mathcal{E}_{\tilde{\varepsilon}_n}^{(NL,p)}(f_n \circ T_n; \tilde{\eta}) \leq C \mathcal{E}_n^{(p)}(f_n; \eta)$$

where  $\mathcal{E}_{\tilde{\varepsilon}_n}^{(NL,p)}$  is defined by

$$(14) \quad \mathcal{E}_{\tilde{\varepsilon}}^{(NL,p)}(f; \eta) = \frac{1}{\varepsilon^p} \int_{\Omega} \int_{\Omega} \eta_{\varepsilon}(|x - z|) |f(x) - f(z)|^p \, dx \, dz.$$

*Proof.* Assume  $\left| \frac{x-z}{\tilde{\varepsilon}_n} \right| < b$  then

$$|T_n(x) - T_n(z)| \leq 2\|T_n - \text{Id}\|_{L^\infty} + |x - z| \leq 2\|T_n - \text{Id}\|_{L^\infty} + b\tilde{\varepsilon}_n = b\varepsilon_n.$$

So,

$$\left| \frac{x-z}{\tilde{\varepsilon}_n} \right| < b \Rightarrow \left| \frac{T_n(x) - T_n(z)}{\varepsilon_n} \right| \leq b$$

and therefore

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

Now,

$$\begin{aligned} \mathcal{E}_{\tilde{\varepsilon}_n}^{(NL,p)}(f_n \circ T_n) &\leq \frac{\varepsilon_n^d}{\tilde{\varepsilon}_n^{d+p}} \int_{\Omega^2} \eta_{\varepsilon_n}(|T_n(x) - T_n(z)|) |f_n(T_n(x)) - f_n(T_n(z))|^p \, dx \, dz \\ &= \frac{\varepsilon_n^{d+p}}{(\inf_{x \in \Omega} \rho^2(x)) \tilde{\varepsilon}_n^{d+p}} \mathcal{E}_n^{(p)}(f_n). \end{aligned}$$

Since  $\frac{\varepsilon_n}{\tilde{\varepsilon}_n} \rightarrow 1$  we are done.  $\square$

In the next lemma we show that that boundedness of non-local energies implies regularity at scales greater  $\varepsilon$ . This allows us to relate non-local bounds to local bounds after mollification.

**Lemma 4.3** (nonlocal to averaged local). *Assume  $\Omega \subset \mathbb{R}^d$  is open and bounded and  $p \geq 1$ . Assume that  $\eta : [0, \infty) \rightarrow [0, \infty)$  is non-increasing,  $\eta(0) > 0$  and  $\eta$  is continuous near 0. Then there exists a constant  $C \geq 1$  and a mollifier  $J$  with  $\text{supp}(J) \subseteq \overline{B(0, 1)}$  such that for all  $\varepsilon > 0$ ,  $f \in L^p(\Omega)$ ,  $\Omega' \subset\subset \Omega$  with  $\text{dist}(\Omega', \partial\Omega) > \varepsilon$  it holds that*

$$\mathcal{E}_\infty^{(p)}(J_\varepsilon * f; \Omega') \leq C \mathcal{E}_\varepsilon^{(NL,p)}(f).$$

where  $\mathcal{E}_\infty^{(p)}$  is defined by (9) and  $\mathcal{E}_\varepsilon^{(NL,p)}$  is defined by (14).

*Proof.* Let  $J$  be a radially symmetric mollifier supported in  $B(0, 1)$  and such that for some  $\beta > 0$ ,  $J \leq \beta\eta$  and  $|\nabla J| \leq \beta\eta$ . Without loss of generality we can assume  $\text{supp}(\eta) \subset \overline{B(0, 1)}$ . Let  $J_\varepsilon(\cdot) = J(\cdot/\varepsilon)/\varepsilon^d$  and let  $g_\varepsilon = J_\varepsilon * f$ . For arbitrary  $x \in \Omega$  with  $\text{dist}(x, \partial\Omega) > \varepsilon$  we have

$$\begin{aligned} |\nabla g_\varepsilon(x)| &= \left| \int_{\Omega} \nabla J_\varepsilon(x-z) f(z) \, dz \right| \\ &= \left| \int_{\Omega} \nabla J_\varepsilon(x-z) (f(z) - f(x)) \, dz - \int_{\mathbb{R}^d \setminus \Omega} \nabla J_\varepsilon(x-z) f(x) \, dz \right| \\ &\leq \frac{\beta}{\varepsilon^{d+1}} \int_{\Omega} \eta \left( \frac{|x-z|}{\varepsilon} \right) |f(z) - f(x)| \, dz + \frac{1}{\varepsilon^{d+1}} \int_{\mathbb{R}^d \setminus \Omega} \left| (\nabla J) \left( \frac{x-z}{\varepsilon} \right) \right| |f(x)| \, dz. \end{aligned}$$

where the second line follows from  $\int_{\mathbb{R}^d} \nabla J(w) \, dw = 0$ . For the second term we have

$$\frac{1}{\varepsilon^{d+1}} \int_{\mathbb{R}^d \setminus \Omega} \left| \nabla J \left( \frac{x-z}{\varepsilon} \right) \right| |f(x)| \, dz = 0$$

since for all  $z \in \mathbb{R}^d \setminus \Omega$  and  $x \in \Omega$  with  $\text{dist}(x, \partial\Omega) > \varepsilon$  it follows that  $|x - z| > \varepsilon$  and thus  $\nabla J\left(\frac{x-z}{\varepsilon}\right) = 0$ . Therefore,

$$\begin{aligned} |\nabla g_\varepsilon(x)|^p &\leq \beta^p \left( \int_{\Omega} \frac{1}{\varepsilon} \eta_\varepsilon(x-z) |f(z) - f(x)| \, dz \right)^p \\ &\leq \gamma_\eta^{p-1} \beta^p \int_{\Omega} \eta_\varepsilon(x-z) \frac{|f(z) - f(x)|^p}{\varepsilon^p} \, dz \end{aligned}$$

by Jensen's inequality and where  $\gamma_\eta = \int_{B(0,1)} \eta(w) \, dw$ . Hence,

$$\begin{aligned} \int_{\Omega'} |\nabla g_\varepsilon(x)|^p \, dx &\leq \gamma_\eta^{p-1} \beta^p \int_{\Omega} \int_{\Omega} \eta_\varepsilon(|x-z|) \left| \frac{f(z) - f(x)}{\varepsilon} \right|^p \, dz \, dx \\ &\leq \gamma_\eta^{p-1} \beta^p \mathcal{E}_\varepsilon^{(NL,p)}(f) \end{aligned}$$

which completes the proof.  $\square$

We prove the compactness property for bounded sequences. The convergence of a subsequence is a consequence of being able to bound  $\tilde{g}_n = J_{\varepsilon_n} * (f_n \circ T_n)$  in  $W^{1,p}$  (hence the sequence  $\{\tilde{g}_n\}_n$  is precompact in  $L^p(\mu)$ ) and show  $\|f_n \circ T_n - \tilde{g}_n\|_{L^p} \rightarrow 0$ .

**Proposition 4.4** (compactness). *Consider the assumptions and the graph construction of Lemma 4.1. Then with probability one, any sequence  $f_n : \Omega_n \rightarrow \mathbb{R}$  with  $\sup_{n \in \mathbb{N}} \mathcal{E}_n^{(p)}(f_n) < \infty$  and  $\sup_{n \in \mathbb{N}} \|f_n\|_{L^\infty(\mu_n)} < \infty$  has a subsequence  $f_{n_m}$  such that  $(\mu_{n_m}, f_{n_m})$ , converges in  $TL^p$  to  $(\mu, f)$  for some  $f \in L^p(\mu)$ .*

*Proof.* Since  $\mathcal{E}_n^{(p)}(f_n) \geq C\mathcal{E}_n^{(1)}(f_n)$  the compactness in  $TL^1$  follows from Theorem 1.2 in [32]. We note that from the proof of Theorem 1.2 it follows that there in fact exists a subsequence  $f_{n_m}$ , and a sequence of transportation maps  $T_{n_m} \# \mu = \mu_{n_m}$  such that

$$\lim_{m \rightarrow \infty} \|f - f_{n_m} \circ T_{n_m}\|_{L^1(\mu)} + \|T_{n_m} - \text{Id}\|_{L^\infty(\mu)} = 0.$$

Since  $\|f - f_{n_m} \circ T_{n_m}\|_{L^\infty(\mu)} \leq M < \infty$  for some  $M \in \mathbb{R}$ , the convergence of  $f_{n_m}$  to  $f$  in  $TL^p$  follows by interpolation.  $\square$

**Lemma 4.5** (uniform convergence). *Consider the assumptions and the graph construction of Lemma 4.1. Assume that  $\varepsilon_n^p n \rightarrow 0$  as  $n \rightarrow \infty$ , which, due to (A5), implies that  $p > d$ . Furthermore assume that with probability one  $(\mu_n, f_n) \rightarrow (\mu, f)$  in  $TL^p$  metric as  $n \rightarrow \infty$  and that  $\sup_{n \in \mathbb{N}} \mathcal{E}_n^{(p)}(f_n) < \infty$ . Then  $f \in C^{0,\gamma}(\Omega)$ , with  $\gamma = 1 - \frac{d}{p} > 0$ , and for all  $\Omega' \subset \subset \Omega$*

$$\max_{\{k : x_k \in \Omega'\}} |f(x_k) - f_n(x_k)| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Moreover, if for all  $k = 1, \dots, N$ ,  $f_n(x_k) = y_k$  for all  $n$ , it follows that  $f(x_k) = y_k$ .

*Proof.* Find constants  $a, b > 0$  such that  $\tilde{\eta}(t) := a$  if  $|t| \leq b$  and  $\tilde{\eta}(t) := 0$  if  $|t| > b$  satisfies  $\tilde{\eta} \leq \eta$ . Now we define  $\tilde{f}_n = f_n \circ T_n$  where  $T_n$  is the transportation map satisfying the conclusions on Theorem 3.3 and set  $\tilde{\varepsilon}_n = \varepsilon_n - \frac{2\|T_n - \text{Id}\|_{L^\infty}}{b}$ . Then for  $n$  sufficiently large  $\tilde{\varepsilon}_n > 0$ , and  $\frac{\varepsilon_n}{\tilde{\varepsilon}_n} \rightarrow 1$ . We note that if  $|T_n(x) - T_n(z)| > b\varepsilon_n$  then

$$|x - z| \geq |T_n(x) - T_n(z)| - 2\|T_n - \text{Id}\|_{L^\infty} > b\varepsilon_n - 2\|T_n - \text{Id}\|_{L^\infty} = \tilde{\varepsilon}_n b.$$

Hence,  $\tilde{\eta} \left( \frac{|x-z|}{\tilde{\varepsilon}_n} \right) \leq \tilde{\eta} \left( \frac{|T_n(x)-T_n(z)|}{\varepsilon_n} \right)$ . Let  $\mathcal{E}_{\tilde{\varepsilon}}^{(NL,p)}$  be the non-local Dirichlet energy defined in (14) with  $\varepsilon = \tilde{\varepsilon}_n$  and  $\eta = \tilde{\eta}$ . Then, by Lemma 4.2

$$\mathcal{E}_{\tilde{\varepsilon}}^{(NL,p)}(\tilde{f}_n) \leq C\mathcal{E}_n^{(p)}(f_n).$$

Hence,  $\mathcal{E}_{\tilde{\varepsilon}}^{(NL,p)}(\tilde{f}_n)$  is bounded and therefore, by Lemma 4.3 we have that  $\mathcal{E}_{\infty}^{(p)}(J_{\tilde{\varepsilon}_n} * f_n; \Omega')$  is bounded for every  $\Omega' \subset\subset \Omega$ . One can easily show  $\|J_{\tilde{\varepsilon}_n} * f_n\|_{L^p(\Omega')} \leq \|\tilde{f}_n\|_{L^p}$  and therefore  $J_{\tilde{\varepsilon}_n} * \tilde{f}_n$  is locally bounded in  $W^{1,p}$ . We also note that since  $f_n \circ T_n$  converges to  $f$  in  $L^p(\mu)$

$$\begin{aligned} \|J_{\tilde{\varepsilon}_n} * \tilde{f}_n - f\|_{L^p(\Omega')} &\leq \|J_{\tilde{\varepsilon}_n} * \tilde{f}_n - J_{\tilde{\varepsilon}_n} * f + J_{\tilde{\varepsilon}_n} * f - f\|_{L^p(\Omega')} \\ &\leq \|\tilde{f}_n - f\|_{L^p(\Omega)} + \|J_{\tilde{\varepsilon}_n} * f - f\|_{L^p(\Omega')} \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Since  $J_{\tilde{\varepsilon}_n} * \tilde{f}_n \rightarrow f$  in  $L^p(\Omega')$ , by the compactness of the embedding of  $W^{1,p}(\Omega')$  into  $C^{0,\gamma}$  (Morrey's inequality), for  $\gamma = 1 - \frac{d}{p}$ , we have that

$$J_{\tilde{\varepsilon}_n} * \tilde{f}_n \rightarrow f \quad \text{uniformly on } \Omega' \text{ as } n \rightarrow \infty.$$

Therefore, for each  $k \in \{1, \dots, N\}$ ,  $J_{\tilde{\varepsilon}_n} * \tilde{f}_n$  converges uniformly to  $f$  on  $B(0, \delta)$  for any  $\delta$  such that  $B(x_k, \delta) \subset \Omega$ . For any  $x \in B(x_k, 3\tilde{\varepsilon}_n) \cap \Omega_n$  we have (for a constant  $C$ )

$$|f_n(x_k) - f_n(x)| \leq \text{osc}_{3\tilde{\varepsilon}_n}(f_n)(x_k) \leq \text{osc}_{4\varepsilon_n}(f_n)(x_k) \leq \left(4^p C \mathcal{E}_n^{(p)}(f_n) n \varepsilon_n^p\right)^{\frac{1}{p}} \rightarrow 0$$

by Lemma 4.1. It follows that

$$\max_{k=1, \dots, n} \max_{x \in B(x_k, 3\tilde{\varepsilon}_n) \cap \Omega_n} |f_n(x) - f_n(x_k)| \rightarrow 0.$$

To complete the proof we notice that for any  $\Omega' \subset\subset \Omega$

$$\begin{aligned} &\max_{\{k : x_k \in \Omega'\}} |f(x_k) - f_n(x_k)| \\ &\leq \max_{\{k : x_k \in \Omega'\}} |f(x_k) - J_{\tilde{\varepsilon}_n} * \tilde{f}_n(x_k)| + |J_{\tilde{\varepsilon}_n} * \tilde{f}_n(x_k) - f_n(x_k)| \\ &\leq \|f - J_{\tilde{\varepsilon}_n} * \tilde{f}_n\|_{L^\infty(\Omega')} + \max_{\{k : x_k \in \Omega'\}} \int_{B(0, 2\tilde{\varepsilon}_n)} J_{\tilde{\varepsilon}_n}(x_k - x) |f_n(T_n(x)) - f_n(x_k)| \, dx \\ &\leq \|f - J_{\tilde{\varepsilon}_n} * \tilde{f}_n\|_{L^\infty(\Omega')} + \max_{\{k : x_k \in \Omega'\}} \sup_{x \in B(x_k, 3\tilde{\varepsilon}_n) \cap \Omega_n} |f_n(x) - f_n(x_k)| \end{aligned}$$

and the above converges to zero for all  $x_k$ . □

#### 4.1 Asymptotic Consistency via $\Gamma$ -Convergence

We approach proving Theorem 2.1 using  $\Gamma$ -convergence. Namely as pointed out in Section 3.1 convergence of minimizers follows from  $\Gamma$ -convergence and compactness. We use the general setup of [32]. In particular we first establish in Lemma 4.6 that nonlocal functionals  $\mathcal{E}_{\varepsilon_n}^{(NL,p)}$   $\Gamma$ -converge to  $\mathcal{E}_{\infty}^{(p)}$ . We then state and prove the  $\Gamma$ -convergence of  $\mathcal{E}_{n, \text{con}}^{(p)}$  towards  $\mathcal{E}_{\infty}^{(p)}$  or  $\mathcal{E}_{\infty, \text{con}}^{(p)}$  depending on how quickly  $\varepsilon_n \rightarrow 0$  as  $n \rightarrow \infty$ . Steps of proving this claim rely on Lemma 4.6.

**Lemma 4.6** (continuum nonlocal to local). *Let  $p > 1$ . Assume  $\Omega$  satisfy the assumptions (A1) - (A2) and  $\eta$  satisfies assumptions (A6) - (A8). Then  $\mathcal{E}_\varepsilon^{(NL,p)}$ , defined in (14),  $\Gamma$ -converges as  $n \rightarrow \infty$  in  $L^p(\Omega)$  to the functional  $\mathcal{E}_\infty^{(p)}$  defined in (9).*

If  $\rho$  is constant and  $\Omega$  is convex this result is contained in the appendix to [3]. For general  $\Omega$  it follows from Theorem 8 in [44]. We remark that while the functional in [44] appears different the term  $|x - y|^p$  which arises can be absorbed in the kernel. The results can be extended to general  $\rho$  in a straightforward manner as has been done for  $p = 1$  in Section 4 of [32] and has been remarked in Proposition 1.10 in [33].

**Theorem 4.7** (discrete to local  $\Gamma$ -convergence). *Let  $p > 1$ . Assume  $\Omega$ ,  $\mu$ ,  $\eta$ ,  $\varepsilon_n$ , and  $x_i$  satisfy the assumptions (A1) - (A8). Let graph weights  $W_{ij}$  be given by (5). Let  $M \geq \max_{i=1,\dots,N} |y_i|$ . Then with probability one  $\mathcal{E}_{n,con}$ , defined in (7),  $\Gamma$ -converges as  $n \rightarrow \infty$  in  $TL^p$  metric on the set  $\{(\nu, g) : \nu \in \mathcal{P}(\Omega), \|g\|_{L^\infty(\nu)} \leq M\}$  to the functional*

$$\begin{cases} \mathcal{E}_{\infty,con}^{(p)} & \text{if } \lim_{n \rightarrow \infty} n\varepsilon_n^p = 0 \\ \mathcal{E}_\infty^{(p)} & \text{if } \lim_{n \rightarrow \infty} n\varepsilon_n^p = \infty \end{cases}$$

where  $\mathcal{E}_\infty^{(p)}$  is defined in (9) and  $\mathcal{E}_{\infty,con}^{(p)}$  is defined in (10).

Restricting the space to the set of functions bounded by  $M$  is really needed only for the case  $\lim_{n \rightarrow \infty} n\varepsilon_n^p = \infty$ . It is required since the functional  $\mathcal{E}_\infty^{(p)}$  is invariant under adding a constant and thus the loss of constraints in the limit when  $\lim_{n \rightarrow \infty} n\varepsilon_n^p = \infty$  would lead to loss of compactness, without the restriction. We note that placing an upper bound on  $f$  is not restrictive in practice since both discrete and continuum minimizers satisfy the bound.

We prove the liminf inequalities and the existence of a recovery sequence separately. Since  $\mathcal{E}_\infty^{(p)} \leq \mathcal{E}_{\infty,con}^{(p)}$  the liminf inequalities needed can be stated in the following way.

**Lemma 4.8.** *Under the same conditions as Theorem 4.7, with probability one, for any  $f \in L^p$  with  $\|f\|_{L^\infty(\mu)} \leq M$  and any sequence  $f_n \rightarrow f$  in  $TL^p$  with  $\|f_n\|_{L^\infty(\mu_n)} \leq M$  we have*

$$(15) \quad \mathcal{E}_\infty^{(p)}(f) \leq \liminf_{n \rightarrow \infty} \mathcal{E}_n^{(p)}(f_n) \leq \liminf_{n \rightarrow \infty} \mathcal{E}_{n,con}^{(p)}(f_n).$$

Furthermore if  $\lim_{n \rightarrow \infty} n\varepsilon_n^p = 0$  then

$$(16) \quad \mathcal{E}_{\infty,con}^{(p)}(f) \leq \liminf_{n \rightarrow \infty} \mathcal{E}_{n,con}^{(p)}(f_n).$$

*Proof.* Let  $f_n \rightarrow f$  in  $TL^p$ . The first inequality of (15) follows from Lemma 4.6 in the same way the analogous result is shown for  $p = 1$  in Section 5 of [32]. The second inequality follows from definition of  $\mathcal{E}_n^{(p)}$  and  $\mathcal{E}_{n,con}^{(p)}$

When  $\lim_{n \rightarrow \infty} n\varepsilon_n^p = 0$  the inequality (16) is a consequence of Lemma 4.5.  $\square$

We now prove the existence of a recovery sequence. Since  $\mathcal{E}_\infty^{(p)} \leq \mathcal{E}_{\infty,con}^{(p)}$  we state it in the following way.

**Lemma 4.9.** *Under the same conditions as Theorem 4.7, with probability one, for any function  $f \in L^p$ , with  $\|f\|_{L^\infty(\mu)} \leq M$  there exists a sequence  $f_n$  satisfying  $f_n \rightarrow f$  in  $TL^p$  with  $\|f_n\|_{L^\infty(\mu_n)} \leq M$  and*

$$(17) \quad \mathcal{E}_{\infty, \text{con}}^{(p)}(f) \geq \limsup_{n \rightarrow \infty} \mathcal{E}_{n, \text{con}}^{(p)}(f_n).$$

Furthermore if  $\lim_{n \rightarrow \infty} n\varepsilon_n^p = \infty$  then

$$(18) \quad \mathcal{E}_{\infty}^{(p)}(f) \geq \limsup_{n \rightarrow \infty} \mathcal{E}_{n, \text{con}}^{(p)}(f_n).$$

*Proof.* The proof of the first inequality is a straightforward adaptation of the analogous result for  $p = 1$  in Section 5 of [32]. The recovery sequence used is defined as a restriction of  $f$  to  $\Omega_n$ :  $f_n(x_i) = f(x_i)$  for all  $i = 1, \dots, n$ , and thus satisfies the constraints and  $\|f_n\|_{L^\infty(\mu_n)} \leq M$ .

The same argument and recovery sequence construction can be used to show that with probability one, for any function  $f \in L^p$ , with  $\|f\|_{L^\infty(\mu)} \leq M$  there exists a sequence  $f_n$  satisfying  $f_n \rightarrow f$  in  $TL^p$  with  $\|f_n\|_{L^\infty(\mu_n)} \leq M$  and

$$(19) \quad \mathcal{E}_{\infty}^{(p)}(f) \geq \limsup_{n \rightarrow \infty} \mathcal{E}_n^{(p)}(f_n).$$

Let us now consider that case that  $n\varepsilon_n^p \rightarrow \infty$  as  $n \rightarrow \infty$  and show the second inequality. Suppose  $\mathcal{E}_{\infty, \text{con}}^{(p)}(f) < \infty$  else the lemma is trivial. Let  $f_n$  be the recovery sequence for (19).

We define  $\hat{f}_n : \Omega_n \rightarrow \mathbb{R}$  by

$$\hat{f}_n(x_i) = \begin{cases} y_i & \text{for } i = 1, \dots, N, \\ f_n(x_i) & \text{for } i = N + 1, \dots, n. \end{cases}$$

We note that  $\hat{f}_n \rightarrow f$  in  $TL^p$  with  $\|\hat{f}_n\|_{L^\infty(\mu_n)} \leq M$ . To show (18) it suffices to show that

$$(20) \quad \lim_{n \rightarrow \infty} \mathcal{E}_n^{(p)}(f_n) - \mathcal{E}_{n, \text{con}}^{(p)}(\hat{f}_n) = 0.$$

We may write,

$$(21) \quad \begin{aligned} \left| \mathcal{E}_n^{(p)}(f_n) - \mathcal{E}_{n, \text{con}}^{(p)}(\hat{f}_n) \right| &\leq \frac{1}{\varepsilon_n^p} \frac{2}{n^2} \sum_{i=1}^N \sum_{j=1}^n \eta_{\varepsilon_n}(|x_i - x_j|) \left| |f(x_i) - f(x_j)|^p - |y_i - f(x_j)|^p \right| \\ &\leq \frac{2^{p+1} M^p}{\varepsilon_n^p n} \sum_{i=1}^N \frac{1}{n} \sum_{j=1}^n \eta_{\varepsilon_n}(|x_i - x_j|) \end{aligned}$$

*Step 1.* Let us consider first the case that  $\eta(t) = a$  if  $|t| < b$  and  $\eta(t) = 0$  otherwise for some  $a, b > 0$ . Then, using Theorem 3.3

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n \eta_{\varepsilon_n}(|x_i - x_j|) &\leq \frac{\eta(0)}{\varepsilon^d} \mu_n(B(x_i, \varepsilon b)) \\ &\leq \frac{\eta(0)}{\varepsilon^d} \mu(B(x_i, \varepsilon b + \|\text{Id} - T_n\|_{L^\infty})) \\ &\leq \eta(0) \left( \frac{\varepsilon b + \|\text{Id} - T_n\|_{L^\infty}}{\varepsilon} \right)^d \text{Vol}(B(0, 1)) \|\rho\|_{L^\infty} \leq C. \end{aligned}$$

Combining this inequality with (21) implies (20).

*Step 2.* Consider now general  $\eta$  satisfying **(A6)**-**(A8)**. Let

$$\tilde{\eta}(t) = \begin{cases} \eta(0) & \text{if } |t| \leq 1 \\ \eta(t) & \text{otherwise.} \end{cases}$$

Note that  $\tilde{\eta}$  is radially nonincreasing,  $\tilde{\eta} \geq \eta$ , and that  $\tilde{\eta}((|x| - 1)_+) \leq \tilde{\eta}(|x|/2)$ . Theorem 3.3 implies that for  $n$  large  $\|\text{Id} - T_n\|_{L^\infty} \leq \varepsilon_n$ . Consequently

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^n \eta_{\varepsilon_n}(|x_i - x_j|) &\leq \frac{1}{n} \sum_{j=1}^n \tilde{\eta}_{\varepsilon_n}(|x_i - x_j|) \\ &= \frac{1}{\varepsilon_n^d} \int_{\Omega} \tilde{\eta} \left( \frac{|x_i - T_n(y)|}{\varepsilon_n} \right) d\mu(y) \\ &\leq \frac{1}{\varepsilon_n^d} \int_{\Omega} \tilde{\eta} \left( \frac{|x_i - y|}{2\varepsilon_n} \right) d\mu(y) \leq C \end{aligned}$$

where the penultimate inequality follows from  $\frac{|x_i - T_n(y)|}{\varepsilon_n} \geq \left( \frac{|x_i - y| - \|T_n - \text{Id}\|_{L^\infty}}{\varepsilon_n} \right)_+ \geq \left( \frac{|x_i - y|}{\varepsilon_n} - 1 \right)_+$ . Again combining this estimate with (21) implies (20).  $\square$

We now state the  $\Gamma$ -convergence result relevant for the penalized model  $\mathcal{S}_n^{(p)}$ .

**Lemma 4.10.** *Under the conditions of Proposition 2.2 we have:*

- (compactness) Any sequence  $f_n : \Omega_n \rightarrow \mathbb{R}$  with  $\sup_{n \in \mathbb{N}} \mathcal{S}_n^{(p)}(f_n) + \|f_n\|_{L^\infty(\mu_n)} < \infty$  has, with probability one, a subsequence  $f_{n_m}$  such that there exists  $f_\infty \in W^{1,p}$  with  $f_{n_m} \rightarrow f_\infty$  in  $TL^p$ .
- ( $\Gamma$ -convergence, well-posed regime) If  $\varepsilon_n^p n \rightarrow 0$  then, with probability one, on the set  $(\mu_n, f_n)$  with  $\|f_n\|_{L^\infty(\mu_n)} \leq M$ ,

$$\Gamma\text{-}\lim_{n \rightarrow \infty} \left( \mathcal{E}_n^{(p)} + \lambda R^{(q)} \right) = \mathcal{E}_\infty^{(p)} + \lambda R^{(q)}$$

where the  $\Gamma$ -convergence is considered in  $TL^p$  topology.

- ( $\Gamma$ -convergence, degenerate regime) If  $\varepsilon_n^p n \rightarrow \infty$  then, with probability one, on the set  $(\mu_n, f_n)$  with  $\|f_n\|_{L^\infty(\mu_n)} \leq M$ ,

$$\Gamma\text{-}\lim_{n \rightarrow \infty} \left( \mathcal{E}_n^{(p)} + \lambda R^{(q)} \right) = \mathcal{E}_\infty^{(p)},$$

where the  $\Gamma$ -convergence is considered in  $TL^p$  topology.

*Proof.* The compactness follows directly from Proposition 4.4.

When  $\varepsilon_n^p n \rightarrow 0$ , for the liminf inequality assume  $f_n \rightarrow f$  in  $TL^p$  and  $\liminf_{n \rightarrow \infty} \mathcal{E}_n^{(p)}(f_n) < \infty$ . Then by Lemma 4.5  $f_n(x_k) \rightarrow f(x_k)$  for all  $k \in \{1, \dots, N\}$  and hence  $\lambda R^{(q)}(f_n) \rightarrow \lambda R^{(q)}(f)$ . By (15) of Lemma 4.8 we have  $\liminf_{n \rightarrow \infty} \left( \mathcal{E}_n^{(p)}(f_n) + \lambda R^{(q)}(f_n) \right) \geq \mathcal{E}_\infty^{(p)}(f) + \lambda R^{(q)}(f)$ . The limsup inequality follows in a similar manner from equation (19) and Lemma 4.5.

If  $\varepsilon_n^p n \rightarrow \infty$ , then the liminf inequality follows from (15) of Lemma 4.8, while, the limsup inequality follows directly from

$$\limsup_{n \rightarrow \infty} \mathcal{E}_n^{(p)}(f_n) + \lambda R^{(q)}(f_n) \leq \limsup_{n \rightarrow \infty} \mathcal{E}_{n, \text{con}}^{(p)}(f_n) \leq \mathcal{E}_\infty^{(p)}(f)$$

and Lemma 4.9.  $\square$

## 4.2 Proofs of Theorem 2.1 and Proposition 2.2

The  $\Gamma$ -convergence and compactness results above allow us to prove Theorem 2.1. It is a general result that  $\Gamma$ -convergence and compactness imply the convergence of minimizers (as well as of almost minimizers) to a minimizer of the limiting problem, see [8, Theorem 1.21] or Theorem 3.2.

*Proof of Theorem 2.1.* Let  $f_n$  be a minimizer of  $\mathcal{E}_{n,con}^{(p)}$ . Recall that  $M \geq \|y\|_{L^\infty(\mu_n)}$ . Note that if  $\|f_n\|_{L^\infty(\mu_n)} > M$  then, since the graph is connected with high probability  $p_n$ , such that  $\sum_{n=1}^{\infty} (1 - p_n) < \infty$ , for  $\hat{f}_n = (f_n \wedge M) \vee (-M)$  we have  $\mathcal{E}_{n,con}^{(p)}(\hat{f}_n) < \mathcal{E}_{n,con}^{(p)}(f_n)$  which contradicts the definition of  $f_n$ . Thus with high probability  $\|f_n\|_{L^\infty} \leq M$  for each  $n$ , hence we can restrict the minimization to the set of  $(f_n, \mu_n)$  such that  $\|f_n\|_{L^\infty(\mu_n)} \leq M$ . This allows us to consider the setting of Theorem 4.7.

By compactness result of Proposition 4.4 there exists a subsequence  $f_{n_m}$  converging in  $TL^p$  to  $f \in L^p(\mu)$ .

To prove (i) assume that  $n\varepsilon_n^p \rightarrow 0$  as  $n \rightarrow \infty$ . The uniform convergence of statement (a) then follows from Lemma 4.5. The  $\Gamma$ -convergence result of Theorem 4.7 implies that  $f$  minimizes  $\mathcal{E}_{\infty,con}^{(p)}$ . Since the minimizer of  $\mathcal{E}_{\infty,con}^{(p)}$  is unique the convergence holds along the whole sequence, thus establishing statement (c).

To prove (ii) assume that  $n\varepsilon_n^p \rightarrow 0$  as  $n \rightarrow \infty$ . Again, Theorem 4.7 implies that  $f$  minimizes  $\mathcal{E}_{\infty}^{(p)}$ .  $\square$

The results of the Proposition 2.2 are proved by the same arguments; using Lemma 4.10 instead of Theorem 4.7.

## 5 Improved Model

In Theorem 2.1 we proved that the model  $\mathcal{E}_{n,con}^{(p)}$ , defined in (7), is consistent as  $n \rightarrow \infty$  and lower bounds (A5) hold, only if

$$\frac{1}{n^p} \gg \varepsilon_n.$$

This upper bound is undesirable as it restricts the range of  $\varepsilon$  that can be used. Furthermore in a nonasymptotic regime, for large but fixed finite  $n$ , it provides no guidance to what  $\varepsilon$  are appropriate (small enough). Finally as our numerical experiments show, see Figures 2(a) and 3(a), the range of  $\varepsilon$  for which the limiting problem is approximated well can be quite narrow. This problem is particularly pronounced if  $p > d$  is close to  $d$ , which is the regime identified in [16] as the most relevant for semi-supervised learning.

It would be advantageous to have another model, asymptotical consistent with  $\mathcal{E}_{\infty,con}^{(p)}$  which would not require an upper bound on  $\varepsilon_n$  (other than  $\varepsilon_n$  to converge to zero) as  $n \rightarrow \infty$ . Here we introduce a new, related, model  $\mathcal{F}_{n,con}^{(p)}$  which has the desired properties, and whose minimizers can be computed with the same algorithms as those for  $\mathcal{E}_{n,con}^{(p)}$ .

We define the set of functions which are constant near the labeled points:

$$C_n^{(\delta)} = \{f : \Omega_n \rightarrow \mathbb{R} : f(x_k) = y_i \text{ whenever } |x_k - x_i| < \delta \text{ for } i = 1, \dots, N\}$$

Let  $L = \min\{|x_i - x_j| : i \neq j\}/2$  and  $R_n = \min\{2\varepsilon_n, L\}$ . The new functional is defined by

$$(22) \quad \mathcal{F}_{n,\text{con}}^{(p)}(f) = \begin{cases} \frac{1}{\varepsilon_n^p} \frac{1}{n^2} \sum_{i,j=1}^n W_{ij} |f(x_i) - f(x_j)|^p & \text{if } f \in C_n^{(R_n)} \\ \infty & \text{else.} \end{cases}$$

We note that for  $f \in C_n^{(R_n)}$ ,  $\mathcal{F}_{n,\text{con}}^{(p)}(f) = \mathcal{E}_{n,\text{con}}^{(p)}(f)$  and that  $\mathcal{F}_{n,\text{con}}^{(p)}(f) \geq \mathcal{E}_{n,\text{con}}^{(p)}(f)$  for all  $f$ .

For the asymptotic consistency we still need to require  $p > d$ , since only then is the limiting model  $\mathcal{E}_{\infty,\text{con}}^{(p)}$  well defined. In Theorem 2.1 this followed from the assumption  $n\varepsilon_n^p \rightarrow 0$  as  $n \rightarrow \infty$ . Since we no longer require the upper bound on  $\varepsilon_n$  we need to require  $p > d$  explicitly.

**Theorem 5.1** (Consistency of the improved model). *Let  $p > d$ . Assume  $\Omega$ ,  $\mu$ ,  $\eta$ , and  $x_i$  satisfy the assumptions (A1) - (A8). Let graph weights  $W_{ij}$  be given by (5). Let  $f_n$  be a sequence of minimizers of  $\mathcal{F}_{n,\text{con}}^{(p)}$  defined in (22). Then, almost surely, the sequence  $(\mu_n, f_n)$  is precompact in the  $TL^p$  metric. The  $TL^p$  limit of any convergent subsequence,  $(\mu_{n_m}, f_{n_m})$ , is of the form  $(\mu, f)$  where  $f \in W^{1,p}(\Omega)$  is a minimizer of  $\mathcal{E}_{\infty,\text{con}}^{(p)}$  defined in (10).*

Proof of the theorem is a straightforward modification of the proof of Theorem 2.1. It relies on the following  $\Gamma$ -convergence result.

**Theorem 5.2** (discrete to local  $\Gamma$ -convergence). *Let  $M \geq \max_{i=1,\dots,N} |y_i|$ . Under the conditions of Theorem 5.1, with probability one  $\mathcal{F}_{n,\text{con}}$   $\Gamma$ -converges as  $n \rightarrow \infty$  in  $TL^p$  metric on the set  $\{(\nu, g) : \nu \in \mathcal{P}(\Omega), \|g\|_{L^\infty(\nu)} \leq M\}$  to the functional  $\mathcal{E}_{\infty,\text{con}}^{(p)}$ .*

We note that statement (15) of Lemma 4.8, and Proposition 4.4 hold for  $\mathcal{F}_{n,\text{con}}^{(p)}$  since  $\mathcal{E}_{n,\text{con}}^{(p)} \leq \mathcal{F}_{n,\text{con}}^{(p)}$ . We now turn to proving the liminf property and the existence of recovery sequence needed to show that  $\mathcal{F}_{n,\text{con}}^{(p)}$   $\Gamma$  converges in  $TL^p$  topology to  $\mathcal{E}_{\infty,\text{con}}^{(p)}$ .

**Lemma 5.3.** *Under the conditions of Theorem 5.1, with probability one, for any  $f \in L^\infty(\mu)$  with  $\|f\|_{L^\infty(\mu)} \leq M$  and any sequence  $f_n \rightarrow f$  in  $TL^p$  with  $\|f_n\|_{L^\infty(\mu_n)} \leq M$  we have*

$$(23) \quad \mathcal{E}_{\infty,\text{con}}^{(p)}(f) \leq \liminf_{n \rightarrow \infty} \mathcal{F}_{n,\text{con}}^{(p)}(f_n).$$

*Proof.* Consider a sequence  $f_n$ , uniformly bounded in  $L^\infty(\mu_n)$  and convergent in  $TL^p$  and such that  $\liminf_{n \rightarrow \infty} \mathcal{F}_{n,\text{con}}^{(p)}(f_n) < \infty$ . Without a loss of generality we assume  $\lim_{n \rightarrow \infty} \mathcal{F}_{n,\text{con}}^{(p)}(f_n) < \infty$ . Note that in contrast to Lemma 4.8 we no longer require  $n\varepsilon_n^p \rightarrow 0$  as  $n \rightarrow \infty$ . Therefore we can no longer use the uniform convergence of Lemma 4.5.

Nevertheless since for  $n$  large  $f_n = y_i$  on  $B(x_i, 2\varepsilon)$  and  $\|\text{Id} - T_n\|_{L^\infty} < \varepsilon$  we conclude that  $\tilde{f}_n := f_n \circ T_n = y_i$  on  $B(x_i, \varepsilon)$  and consequently that for  $g_n := J_{\varepsilon_n} * \tilde{f}_n$  it holds that  $g_n(x_i) = y_i$ . Furthermore note that  $\|g_n\|_{L^\infty} \leq M$ . By bounds of Lemma 4.2 and Lemma 4.3,  $g_n$  is uniformly bounded in  $W^{1,p}(\Omega')$  for any  $\Omega' \subset\subset \Omega$ . Arguing as in the proof of Lemma 4.5 we conclude that  $g_n \rightarrow f$  in  $L^p(\Omega)$ . Since  $p > d$ ,  $W^{1,p}$  is compactly embedded in the space of continuous functions. This implies that  $g_n$  uniformly converges to  $f$  on sets compactly contained in  $\Omega$ . Therefore  $f(x_i) = y_i$  for all  $i = 1, \dots, N$ . Combining this with statement (15) of Lemma 4.8 yields (23).  $\square$

**Lemma 5.4.** *Under the conditions of Theorem 5.1, with probability one, for any  $f \in L^\infty(\mu)$  with  $\|f\|_{L^\infty(\mu)} \leq M$  there exists a sequence  $f_n \rightarrow f$  in  $TL^p$  with  $\|f_n\|_{L^\infty(\mu_n)} \leq M$  such that*

$$(24) \quad \mathcal{E}_{\infty,\text{con}}^{(p)}(f) \geq \limsup_{n \rightarrow \infty} \mathcal{F}_{n,\text{con}}^{(p)}(f_n).$$

*Proof.* Assume  $\|f\|_{L^\infty(\mu)} \leq M$  and  $\mathcal{E}_{\infty, \text{con}}^{(p)}(f) < \infty$ . Then  $f \in W^{1,p}(\Omega)$  and since  $p > d$ ,  $f$  is continuous. Furthermore  $f(x_i) = y_i$  for all  $i = 1, \dots, N$ .

If there exists  $\delta > 0$  such that  $f \in W^{1,p}(\Omega)$  satisfies  $f(x) = y_i$  for all  $x \in B(x_i, \delta)$  and  $i = 1, \dots, N$  then the proof of (24) is the same as the proof of (17). In particular one can use the restriction of  $f$  to data points to construct a recovery sequence.

To treat general  $f$  in  $W^{1,p}(\Omega)$  it suffices to find a sequence  $g_n \in W^{1,p}(\Omega)$  satisfying the conditions above, namely such that  $\|g_n\|_{L^\infty} \leq M$ ,  $g_n(x) = y_i$  for all  $x \in B(x_i, \delta_n)$  for a sequence  $\delta_n \geq R$  converging to zero, which satisfies

$$(25) \quad \lim_{n \rightarrow \infty} \mathcal{E}_{\infty, \text{con}}^{(p)}(g_n) = \mathcal{E}_{\infty, \text{con}}^{(p)}(f).$$

We construct the sequence in the following way. Let  $\theta$  be a cut-off function supported in  $B(0, 2)$ . That is assume  $\theta : \mathbb{R}^d \rightarrow [0, 1]$  is smooth, radially symmetric and nonincreasing such that  $\theta = 1$  on  $B(0, 1)$ ,  $\theta = 0$  outside of  $B(0, 2)$ , and  $|\nabla\theta| < 2$ . Define  $\theta_\delta(z) = \theta(z/\delta)$ .

We first consider the case  $N = 1$ . Let

$$g_n(x) = (1 - \theta_{\delta_n}(x - x_1))f(x) + \theta_{\delta_n}(x - x_1)y_1.$$

Then

$$\left| \mathcal{E}_{\infty, \text{con}}^{(p)}(g_n) - \mathcal{E}_{\infty, \text{con}}^{(p)}(f) \right| \geq \sigma_n \int_{\Omega} \left| |\nabla g_n|^p - |\nabla f|^p \right| \rho^2 dx \leq \sigma_n \int_{B(0, 2\delta_n)} (|\nabla g_n|^p + |\nabla f|^p) \rho^2 dx$$

We estimate

$$\int_{B(0, 2\delta_n)} |\nabla g_n|^p \rho^2 dx \leq 2^p \int_{B(0, 2\delta_n)} |(f(x_1) - f(x))\nabla\theta_{\delta_n}(x - x_1)|^p + |\nabla f(x)|^p \rho^2 dx$$

Using that  $f \in C^{0, 1-d/p}$  and furthermore, by the remark following Theorem 4 in Section 5.6.2 of [20] we obtain

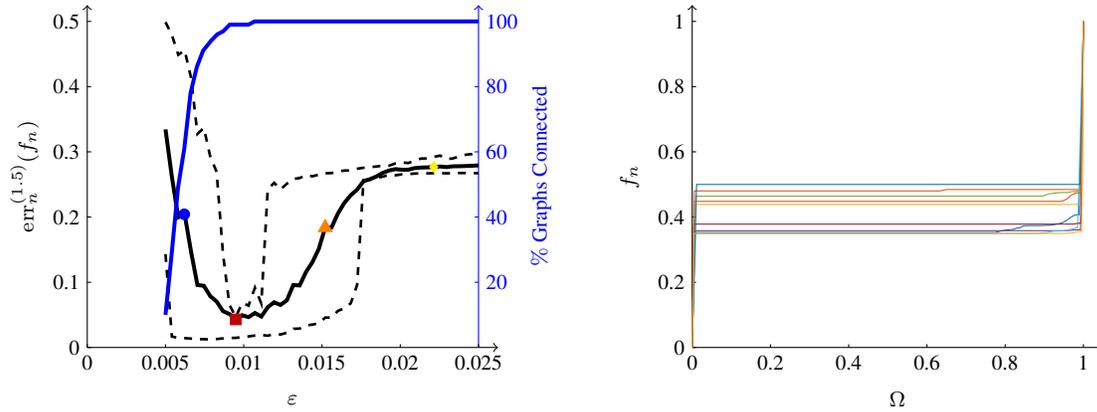
$$\begin{aligned} \int_{B(0, 2\delta_n)} |(f(x) - f(x_1))\nabla\theta_{\delta_n}(x - x_1)|^p \rho^2(x) dx &\leq C_1 \delta_n^{p-d} \|\nabla f\|_{L^p(B(0, 2\delta_n))}^p \|\nabla\theta_{\delta_n}\|_{L^p(B(x_1, 2\delta_n))}^p \\ &\leq C_1 \|\nabla f\|_{L^p(B(0, 2\delta_n))}^p \|\nabla\theta\|_{L^p(\mathbb{R}^d)}^p. \end{aligned}$$

Since  $\lim_{n \rightarrow \infty} \int_{B(0, 4\delta_n)} |\nabla f(x)|^p dx = 0$ , by combining the inequalities above we conclude that (25) holds.

Generalizing to  $N > 1$  is straightforward.  $\square$

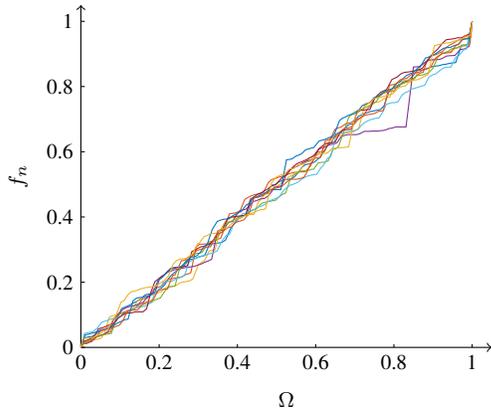
## 6 Numerical Experiments

The results of Theorem 2.1 show that when  $\varepsilon_n^p n \rightarrow 0$  then the solutions to the SSL problem (7) converge to a solution of the continuum constrained problem (9), while when  $\varepsilon_n^p n \rightarrow \infty$  they degenerate as  $n \rightarrow \infty$ . However, in practice, for finite  $n$ , this does not provide a precise guidance on what  $\varepsilon$  are appropriate. We investigate, via numerical experiments in 1D and 2D, the affect of  $\varepsilon$  on solutions to (7) in elementary examples. We also numerically compare the results with our improved model (22).

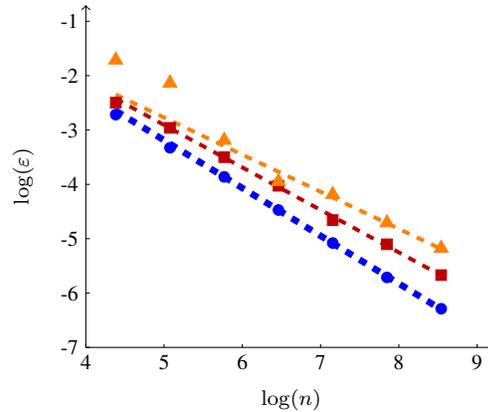


(a) Error (26) for  $n = 1280$ . Black line is the mean error, dashed lines are the 10% and 90% quantiles. Connectivity bound  $\varepsilon_{\text{conn}}$  in blue, optimal  $\varepsilon_*^{(1.5)}$  in red and upper bound  $\varepsilon_{\text{upper}}^{(1.5)}$  in orange. Blue line is the percentage of connected graphs for a given  $\varepsilon$ .

(b) We plot the functions output by the algorithm corresponding to nine realizations of the data for  $n = 1280$  and  $\varepsilon = 0.022$  (marked in yellow in Figure (a)).



(c) The output of the algorithm,  $f_n$ , for nine realizations of the data for  $n = 1280$  and  $\varepsilon = \varepsilon_*^{(1.5)}$ .



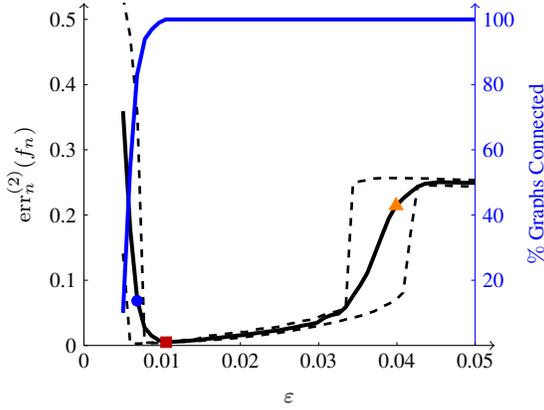
(d) Orange triangles dots are  $\varepsilon_{\text{upper}}^{(1.5)}$ , red squares are  $\varepsilon_*^{(1.5)}$ , and blue dots are  $\varepsilon_{\text{conn}}$ . Lines show the linear fit over last 5 points.

Figure 2: 1D numerical experiments for (7) with  $p = 1.5$ , averaged over 100 realizations.

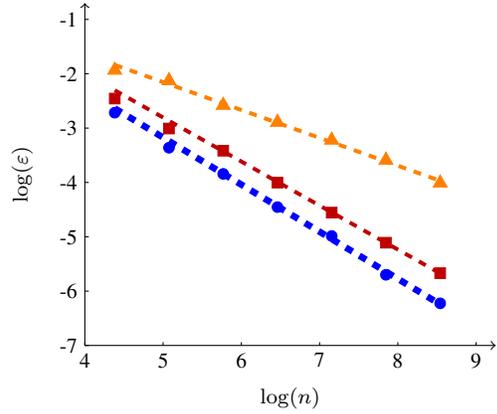
## 6.1 1D Numerical Experiments

Let  $\mu$  be the uniform measure on  $[0, 1]$  and consider  $\eta$  defined by  $\eta(t) = 1$  if  $t \leq 1$  and  $\eta(t) = 0$  otherwise. We consider two different values of  $p$ :  $p = 1.5$  and  $p = 2$ . The training set is  $\{(0, 0), (1, 1)\}$ , that is we condition on functions  $f_n$  taking the value 0 at  $x_1 = 0$  and taking the value 1 at  $x_2 = 1$  (so  $N = 2$ ). We avoid using  $p = 1$  since any increasing function  $f$  with  $f(0) = 0$  and  $f(1) = 1$  is a minimizer to the limiting problem. For  $p > 1$  the solution to the constrained limiting problem is  $f^\dagger(x) = x$  (note that this is independent of  $p$ ). Since  $f^\dagger$  is continuous we can consider the following simple-to-compute notion of error:

$$(26) \quad \text{err}_n^{(p)}(f_n) = \|f_n - f^\dagger\|_{L^p(\mu_n)}.$$

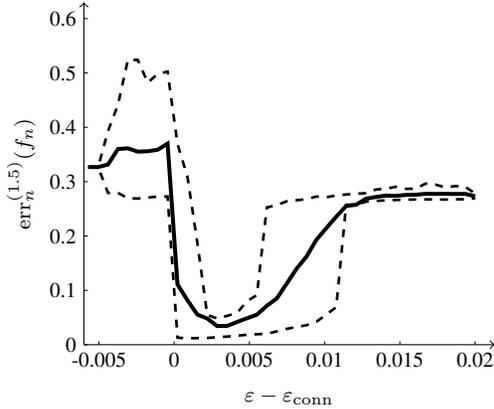


(a) Error (26) for  $n = 1280$ . Black line is the mean error, dashed lines are the 10% and 90% quantiles. Connectivity bound  $\epsilon_{\text{conn}}$  in blue, optimal  $\epsilon_*^{(2)}$  in red and upper bound  $\epsilon_{\text{upper}}^{(2)}$  in orange. Blue line is the percentage of connected graphs for a given  $\epsilon$ .

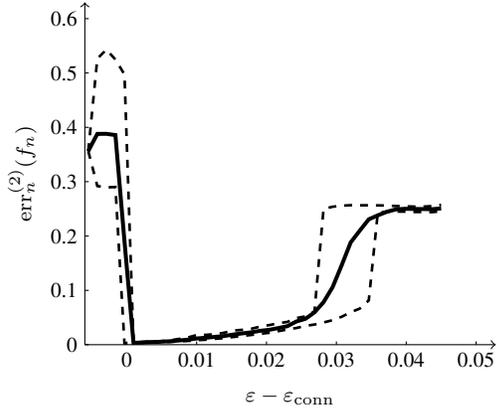


(b) Orange triangles dots are  $\epsilon_{\text{upper}}^{(2)}$ , red squares are  $\epsilon_*^{(2)}$ , and blue dots are  $\epsilon_{\text{conn}}$ . Lines show the linear fit over last 5 points.

Figure 3: 1D numerical experiments averaged over 100 realizations for (7) with  $p = 2$ .



(a) Error (26) for the output  $f_n$  for  $n = 1280$  and  $p = 1.5$ . The solid line is the mean error, the dashed lines are the 10% and 90% quantiles.

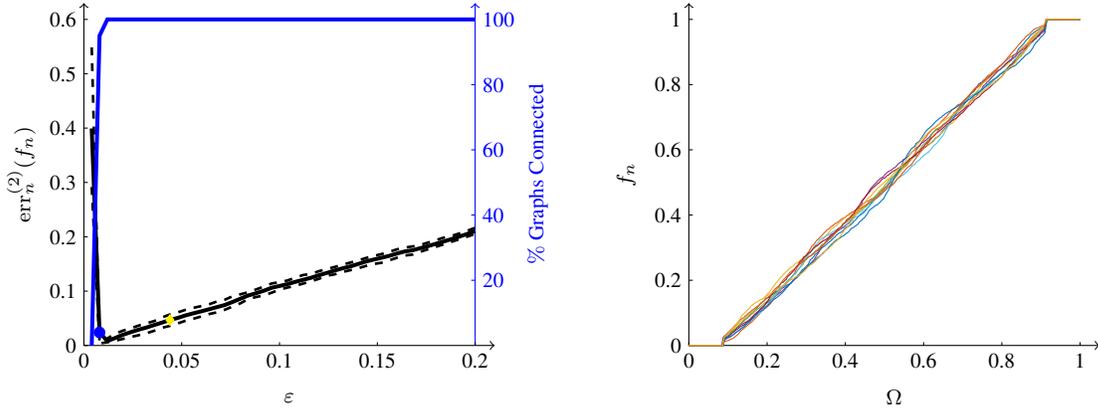


(b) Error (26) for the output  $f_n$  for  $n = 1280$  and  $p = 2$ . The solid line is the mean error, the dashed lines are the 10% and 90% quantiles.

Figure 4: Error shifted by connectivity radius using the same results as in Figures 2 and 3.

To find minimizers of (7) we use coordinate gradient descent. The number of data points varies from  $n = 80$  to  $n = 5120$ . For each  $n$ ,  $\epsilon$  and  $p$  we consider 100 different realizations of the random sample and plot the average results. When  $\epsilon$  is too small the graph is disconnected and we should not expect informative solutions, when  $\epsilon$  is large we expect discontinuities to arise and cause degeneracy. In Figure 2(a) and Figure 3(a) we plot the error as a function of  $\epsilon$  for fixed  $n = 1280$ . We see clear regions where  $\epsilon$  is too small and where  $\epsilon$  is too large, with the intermediate range producing good estimators. Plots of minimizers for a particular  $\epsilon$  in the “large- $\epsilon$ ” region, show that they exhibit discontinuities, as expected.

To measure how the transition point in  $\epsilon$  where minimizers change behavior scale with  $n$  we define



(a) Error (26) for  $n = 1280$ . Black line is the mean error, dashed lines are the 10% and 90% quantiles. Blue dot is the connectivity bound  $\epsilon_{\text{conn}}$ . Blue line is the percentage of connected graphs for given  $\epsilon$ .

(b) We plot the functions output from the algorithm corresponding to multiple realizations of the data for  $n = 1280$  and  $\epsilon = 0.045$  (marked in yellow in Figure (a)).

Figure 5: 1D numerical experiments averaged over 100 realizations for model (22) with  $p = 2$ .

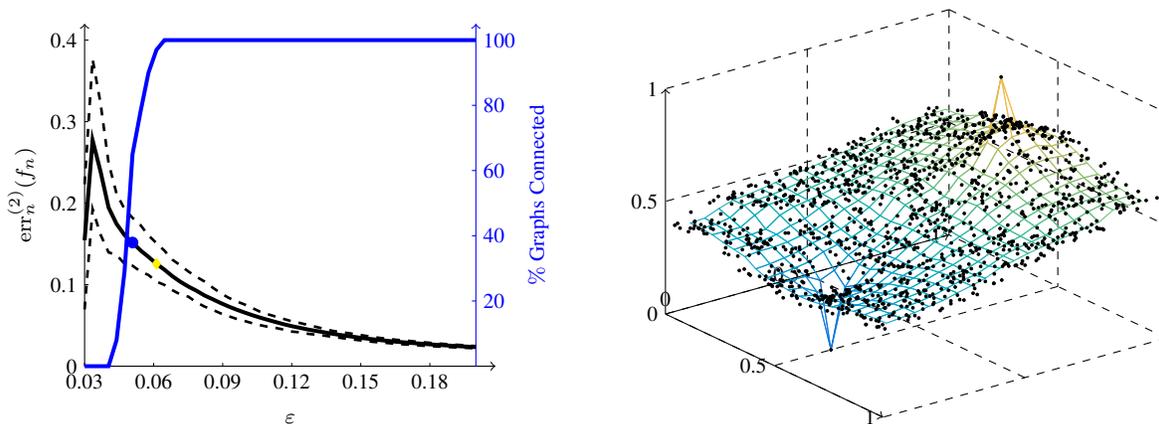
the following:

- (i) Given a realization  $\{x_i^\omega\}_{i=1}^n$  let  $\epsilon_{\text{conn}}(n; \omega)$  be the connectivity radius for the particular realization,  $\omega$  that is the smallest  $\epsilon$  such that the graph with weights  $W_{ij} = \eta_\epsilon(|x_i - x_j|)$  is connected. The value  $\epsilon_{\text{conn}}(n) = \frac{1}{M} \sum_{i=1}^M \epsilon_{\text{conn}}(n; \omega_i)$  is the connectivity radius averaged over the realizations considered. We considered  $M = 100$  realizations.
- (ii)  $\epsilon_*^{(p)}(n)$  is the empirically best choice for  $\epsilon$ , namely the  $\epsilon$  that minimizes  $\text{err}_n^{(p)}(f_n)$  where  $f_n$  is the minimizer of (7) with  $\epsilon_n = \epsilon$ ; again averaged over  $M = 100$  realizations.
- (iii)  $\epsilon_{\text{upper}}^{(p)}(n)$  is the upper bound on  $\epsilon$  for which the algorithm behaves well, which we identify as the maximizer of the second derivative of  $-\text{err}_n^{(p)}(f_n)$  with respect to  $\epsilon$ , among  $\epsilon \geq \epsilon_*^{(p)}(n)$ . While computing  $\epsilon_{\text{upper}}^{(p)}(n)$  we smooth the error slightly so that the method is robust to small perturbations. As above the value is averaged over 100 realizations.

All of these points are highlighted in Figure 2(a) and Figure 3(a). In Figure 2(d) and Figure 3(d) we plot how these values of  $\epsilon$  scale with  $n$ . The best linear fit (based on five largest values of  $n$ ) in the log-log domain gives the following scalings

$$\begin{aligned}
 \epsilon_*^{(1.5)} &\approx \frac{2.719}{n^{0.781}} & \epsilon_{\text{upper}}^{(1.5)} &\approx \frac{1.905}{n^{0.683}} \\
 \epsilon_*^{(2)} &\approx \frac{3.472}{n^{0.810}} & \epsilon_{\text{upper}}^{(2)} &\approx \frac{1.507}{n^{0.513}} \\
 \epsilon_{\text{conn}} &\approx \frac{3.342}{n^{0.879}}.
 \end{aligned}$$

We observe that asymptotic scaling established in Theorem 2.1 for  $\epsilon_{\text{upper}}^{(p)}$  is  $\frac{1}{n^{0.5}}$  for  $p = 2$  and  $\frac{1}{n^{0.667}}$  for  $p = 1.5$ , which is very close to our numerical results. The true scaling in the connectivity of the graph is  $\frac{\log(n)}{n}$ , our numerical results behave approximately as  $\frac{1}{n^{0.879}}$ . We note that if we use a linear fit in the log-log domain over the same range as considered above then  $\frac{n}{\log(n)}$  is approximated by  $\frac{1}{n^{0.859}}$ .



(a) Error (28) for  $n = 1280$ . Black line is the mean error, dashed lines are the 10% and 90% quantiles. Blue dot is the connectivity bound  $\varepsilon_{\text{conn}}$ . Blue line is the percentage of connected graphs for given  $\varepsilon$ .

(b) We plot an example of a function output from the algorithm corresponding to  $n = 1280$  and  $\varepsilon = 0.06$  (marked in yellow in Figure (a)). The grid is to aid visualisation.

Figure 6: 2D numerical experiments averaged over 100 realizations for (7) with  $p = 2$ .

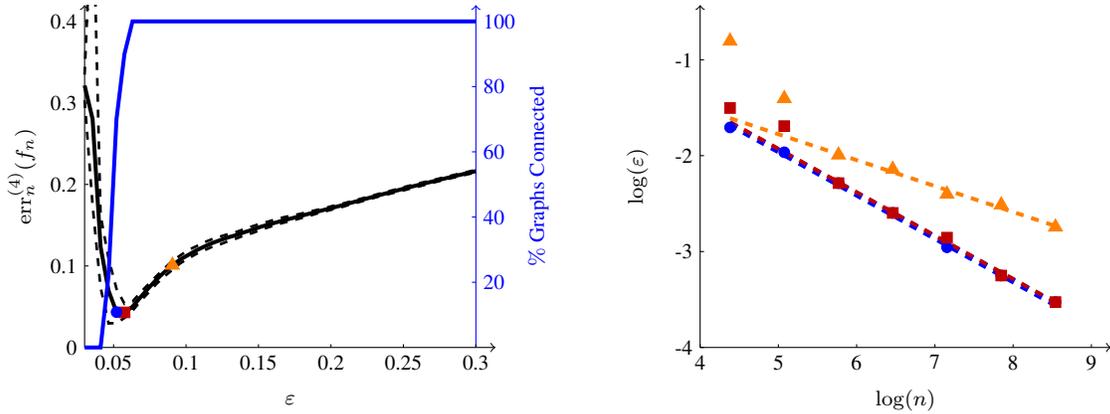
We observe that optimal choice  $\varepsilon_*^{(p)}$  is quite close to the connectivity radius  $\varepsilon - \varepsilon_{\text{conn}}(n)$ . Choosing  $\varepsilon$  smaller than this results in a big error from a small number of realizations. To further investigate the proximity of the connectivity radius and the optimal choice of  $\varepsilon$  we plot in Figure 4 the error as the function of the size of  $\varepsilon$  relative to the connectivity radius. More precisely we consider  $\text{err}_n^{(p)}(f_n; \varepsilon, \omega)$ , where  $f_n$  is the minimizer of (7) for given  $\varepsilon$  and realization  $\omega$ , as a function of  $\varepsilon - \varepsilon_{\text{conn}}(n; \omega)$  and then average over  $M = 100$  realizations. We observe that, for both  $p = 1.5$  and  $p = 2$ , the error is the smallest when  $\varepsilon$  is quite close to the connectivity radius. The slight difference is that for  $p = 1.5$  there is a short interval beyond the connectivity radius where the error is still decreasing.

*Remark 6.1.* The close proximity of the optimal epsilon to the connectivity radius, both for the original model and the improved model (Figure 5[a]) and both in 1D and 2D (Figure 7[a]), is not obvious since for  $\varepsilon$  small (i.e. relatively close to the connectivity radius)  $\mathcal{E}_n^{(p)}(f)$  is a poor approximation of  $\mathcal{E}_\infty^{(p)}(f)$ , even for  $f$  a fixed smooth function. Explaining the observed behavior of the error is an interesting open problem, that we believe should be approached from the viewpoint to stochastic homogenization.

The improved model (22), for which we show results in Figure 5, is far more robust to the choice of  $\varepsilon$ . We plot the error as a function of  $\varepsilon$  for  $n = 1280$  and we see a much larger range in the admissible choices of  $\varepsilon$ . To highlight the difference we plot in Figure 5(b) outputs from multiple realizations of the data under the same conditions as for Figure 3(b), in particular we use the same choice of  $\varepsilon$ . Note that the horizontal axis covers a much larger range on Figure 5(b). The comparison shows that model (7) does not produce a reasonable output when  $\varepsilon \gtrsim 0.04$ , while all outputs of (22) (when  $\varepsilon$  is larger than the connectivity radius) are close to the truth.

## 6.2 2D Numerical Experiments

Let  $\mu$  be the uniform measure on  $\Omega = [0, 1] \times [0, 1]$ , and  $\eta(t) = 1$  if  $|t| \leq 1$ ,  $\eta(t) = 0$  otherwise. In 2D the critical value of  $p$  is  $p = 2$ , and we therefore choose to investigate  $p = 2$  and  $p = 4$ . The training set is  $x_1 = (0.2, 0.5)$ ,  $x_2 = (0.8, 0.5)$ , with labels  $y_1 = 0$ ,  $y_2 = 1$ . In contrast to the 1D



(a) Error (27) for  $n = 1280$ . Black line is the mean error, dashed lines are the 10% and 90% quantiles. Blue dot is the connectivity bound  $\varepsilon_{\text{conn}}$ , red is  $\varepsilon_*^{(4)}$ , and orange is the upper bound  $\varepsilon_{\text{upper}}^{(4)}$ . Blue line is the percentage of connected graphs for given  $\varepsilon$ .

(b) Orange triangles are  $\varepsilon_{\text{upper}}^{(4)}$ , red squares are  $\varepsilon_*^{(4)}$ , and blue dots are  $\varepsilon_{\text{conn}}$ . Lines show the linear fit over last 5 points.

Figure 7: 2D numerical experiments averaged over 100 realizations for (7) with  $p = 4$ .

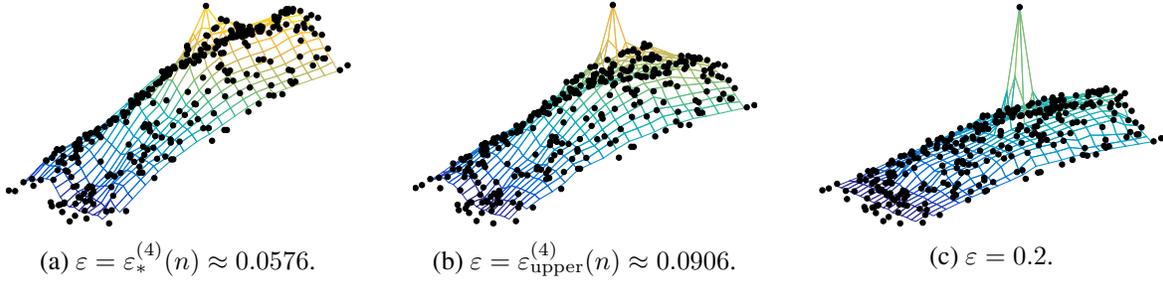


Figure 8: Realizations of (7) with  $p = 4$  and  $n = 1280$  for a select choices of  $\varepsilon$ . Only the part of the domain near labeled point  $x_2$  is shown. The grids are to aid visualization.

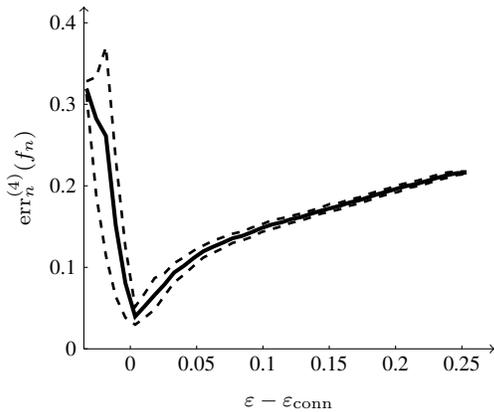
example the solution to the continuum problem (10) (in the well-posed regime) depends on  $p$  and furthermore cannot be solved analytically. To estimate the solution we discretised (10) on a uniform grid and ran a gradient descent algorithm to approximate the minimiser. In the case when  $p = 4$  we plot our numerical approximation of the continuum minimizer to (10) in Figure 1(b). For  $p > 2$  we define the error by

$$(27) \quad \text{err}_n^{(p)}(f_n) = \|f_n - f^{p,\dagger}\|_{L^p(\mu_n)}$$

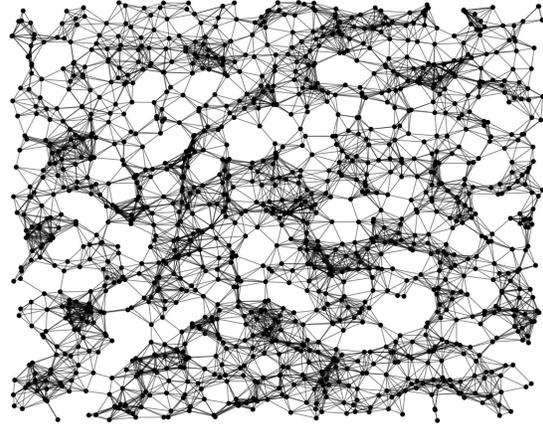
where  $f^{p,\dagger}$  minimizes (10). In the ill-posed case ( $p \leq 2$ ) any constant function is a minimizer to the continuum problem, in which case we define the error as

$$(28) \quad \text{err}_n^{(p)}(f_n) = \inf_{c \in \mathbb{R}} \|f_n - c\|_{L^p(\mu_n)}.$$

To find minimizers of (7) for  $p = 4$  we use coordinate gradient descent. For  $p = 2$  we use the method of [61] that exactly solves the Euler Lagrange equation  $((L_n f)_i = 0$ , where  $L_n$  is the graph



(a) Error (27) using the same results as in Figure 7. The solid line is the mean error, the dashed lines are the 10% and 90% quantiles.



(b) Example graph in 2D for  $\varepsilon = \varepsilon_*^{(4)}(n)$  and  $n = 1280$ .

Figure 9: Error dependency on the connectivity radius and the graph for optimal  $\varepsilon$ , for  $n = 1280$  and  $p = 4$ .

Laplacian, for  $i > 2$  and  $f_1 = 0, f_2 = 1$ ). The number of data points varies from 80 to 5120. We use 100 different realizations of the data  $\{x_i\}_{i=1}^n$  for each  $n$  and each choice of  $\varepsilon$ .

Figure 6 shows the results in the ill-posed regime, for  $p = 2$  and  $n = 1280$ . We observe that the solutions form spikes in order to satisfy the constraints. Spikes are present for all  $\varepsilon$  beyond the connectivity threshold, and grow as  $\varepsilon$  increases (recall that the solution to the continuum problem is a constant and therefore the error decreasing indicates convergence to a constant solution with spikes around the two training data points).

In Figures 1 and 7 show the results for  $p = 4$  which is in the well-posed range. Figure 1(a) presents the numerically computed discrete minimizer for the optimal radius  $\varepsilon = \varepsilon_*^{(4)}$ . We observe in Figure 7(a) that, similarly to 1D, for  $\varepsilon$  below the average connectivity radius, or  $\varepsilon$  is large, the error is high, and is the lowest for  $\varepsilon$  in between and close to connectivity threshold. In contrast to the 1D results we do notice that the transitions between the well-posed and ill-posed regime is gradual.

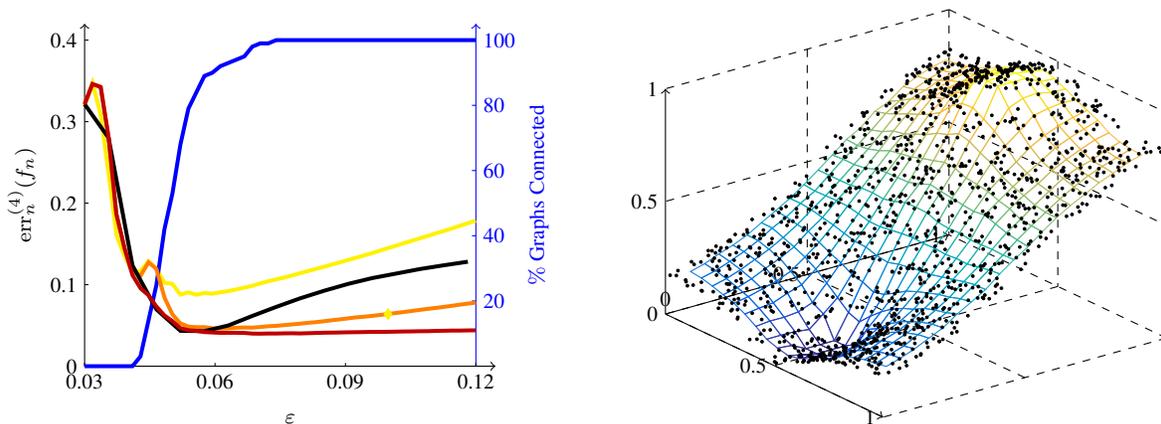
The numerical scaling for  $\varepsilon_*^{(p)}$ ,  $\varepsilon_{\text{upper}}^{(p)}$ , and  $\varepsilon_{\text{conn}}$  (with the same definitions of quantities as in the 1D experiments in the previous subsection) we find is

$$\varepsilon_*^{(4)} \approx \frac{1.394}{n^{0.452}} \quad \varepsilon_{\text{upper}}^{(4)} \approx \frac{0.654}{n^{0.270}} \quad \varepsilon_{\text{conn}} \approx \frac{1.368}{n^{0.452}}.$$

The connectivity radius should scale according to  $\sqrt{\frac{\log(n)}{n}}$ , which is close to our observed rate of  $n^{-0.452}$  (in fact when linear fitting  $\log \varepsilon$  to  $\frac{\log n}{n}$  one obtains  $\varepsilon_{\text{conn}} \approx 0.829 \left(\frac{\log n}{n}\right)^{0.526}$ ). Our theoretical predictions give the scaling of the upper bound as  $\varepsilon_{\text{upper}}^{(4)} \asymp n^{-0.25}$ , close to our numerical rate of  $n^{-0.270}$ .

In Figure 8 we show instances of numerically computed minimizers of (7) for increasing values of  $\varepsilon$ . They show that the breakdown of the numerical approximation of the continuum solution (shown in Figure 10(c)) happens via development of spikes.

As in the 1D examples (shown on Figure 4) we investigate the proximity of the optimal radius  $\varepsilon_*^{(p)}(n; \omega)$  to the connectivity radius  $\varepsilon_{\text{conn}}(n; \omega)$ , where  $\omega$  is the sample considered. In Figure 9 we



(a) The black line is the mean error (27) for model (7). The error for the improved model (22) with constraint radius  $R_n = 2\epsilon$  is in yellow,  $R_n = \epsilon$  is in orange, and  $R_n = \epsilon/2$  is in red.

(b) We plot an example of a function output from the algorithm corresponding to  $n = 1280$  and  $\epsilon = 0.1$  for the constraint set of size  $\epsilon$  (marked in yellow in Figure (a)). The grid is to aid visualisation.

Figure 10: Experiments for improved model (22) with  $n = 1280$  and  $p = 4$  averaged over 100 realizations.

plot the error,  $\text{err}_n^{(p)}(f_n; \omega)$ , against  $\epsilon - \epsilon_{\text{conn}}(n; \omega)$  for  $n = 1280$  and  $p = 4$ , averaged over 100 samples. The phenomena we observe is similar to the 1D case; the error is large and highly variable for  $\epsilon$  below the connectivity radius. There is a sharp transition to the well-posed regime, as soon as the graph is connected with the error increasing with  $\epsilon$ . As we explain in Remark 6.1 it is an intriguing and important open problem to explain why the error is the smallest for rather coarse graphs (Figure 9(b)).

Our theoretical result in Section 5 showed that minimizers of the improved model, (22), converge as  $n \rightarrow \infty$  to the correct solution if  $1 \gg \epsilon_n \gg (\ln n/n)^{1/d}$ , regardless of how slowly  $\epsilon_n \rightarrow 0$ . Here we numerically investigate two issues. One is how precisely does the error of the improved model depend on  $\epsilon$  for fixed  $n$ . The other is to compare the observed error of the improved model to the original model. Recall that in the improved model we prove convergence when the labels are extended around the training set to balls of radius  $2\epsilon$ . This is needed in our proof to ensure that spikes do not form. Here we numerically investigate if extending the labels to smaller balls is sufficient to prevent the spike formation. In particular in Figure 10(a) we display the error for fixed  $n = 1280$  and constraint ball radii  $2\epsilon$ ,  $\epsilon$  and  $\epsilon/2$ . The numerics show that even radius  $\epsilon/2$  is sufficient to prevent spike formation and that it allows for better approximation of the continuum solution. We also observe that fixing the labels on larger sets can significantly impact the accuracy of approximation. This issue is less pronounced for larger values of  $n$ , where the connectivity radius is small compared to distances between the labeled points.

## Acknowledgements

The authors thank Matt Dunlop and Andrew Stuart for enlightening exchanges. The authors are grateful to Nicolás García Trillos for careful reading of the manuscript and insightful remarks. This material is based on work supported by the National Science Foundation under the grants CCT 1421502 and DMS 1516677. The authors are also grateful to the Center for Nonlinear Analysis (CNA) for support. MT is grateful for the support of the Cantab Capital Institute for the Mathematics of Information at the University of Cambridge.

## References

- [1] M. Ajtai, J. Komlós, and G. Tusnády. On optimal matchings. *Combinatorica*, 4(4):259–264, 1984.
- [2] M. Alamgir and U. Von Luxburg. Phase transition in the family of  $p$ -resistances. In *Advances in Neural Information Processing Systems (NIPS)*, pages 379–387, 2011.
- [3] G. Alberti and G. Bellettini. A non-local anisotropic model for phase transitions: asymptotic behaviour of rescaled energies. *European J. Appl. Math.*, 9(3):261–284, 1998.
- [4] M. Belkin and P. Niyogi. Using manifold structure for partially labeled classification. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.
- [5] M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine learning*, 56(1):209–239, 2004.
- [6] M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems (NIPS)*, pages 129–136, 2007.
- [7] Andrea L Bertozzi, Xiyang Luo, Andrew M Stuart, and Konstantinos C Zygalakis. Uncertainty quantification in the classification of high dimensional data. *arXiv preprint arXiv:1703.08816*, 2017.
- [8] A. Braides.  *$\Gamma$ -Convergence for Beginners*. Oxford University Press, 2002.
- [9] T. Bühler and M. Hein. Spectral clustering based on the graph  $p$ -Laplacian. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 81–88, 2009.
- [10] D. Burago, S. Ivanov, and Y. Kurylev. A graph discretization of the Laplace-Beltrami operator. *J. Spectr. Theory*, 4(4):675–714, 2014.
- [11] J. Calder, S. Esedoǧlu, and A. O. Hero. A Hamilton-Jacobi equation for the continuum limit of nondominated sorting. *SIAM J. Math. Anal.*, 46(1):603–638, 2014.
- [12] R. R. Coifman and S. Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.
- [13] M. G. Crandall, L. C. Evans, and R. F. Gariepy. Optimal Lipschitz extensions and the infinity Laplacian. *Calculus of Variations and Partial Differential Equations*, 13(2):123–139, 2001.
- [14] G. Dal Maso. *An Introduction to  $\Gamma$ -Convergence*. Springer, 1993.
- [15] E. Davis and S. Sethuraman. Consistency of modularity clustering on random geometric graphs. *arXiv preprint arXiv:1604.03993*, 2016.
- [16] 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*, pages 879–906, 2016.
- [17] A. Elmoataz, X. Desquesnes, and O. Lezoray. Non-local morphological PDEs and  $p$ -Laplacian equation on graphs with applications in image processing and machine learning. *IEEE Journal of Selected Topics in Signal Processing*, 6(7):764–779, 2012.

- [18] A. Elmoataz, F. Lozes, and M. Toutain. Nonlocal pdes on graphs: From tug-of-war games to unified interpolation on images and point clouds. *Journal of Mathematical Imaging and Vision*, 57(3):381–401, 2017.
- [19] A. Elmoataz, M. Toutain, and D. Tenbrinck. On the  $p$ -Laplacian and  $\infty$ -Laplacian on graphs with applications in image and data processing. *SIAM Journal on Imaging Sciences*, 8(4):2412–2451, 2015.
- [20] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.
- [21] C. Fefferman. Fitting a  $C^m$ -smooth function to data III. *Annals of Mathematics*, 2009.
- [22] C. Fefferman, A. Israel, and G. K. Luli. Fitting a Sobolev function to data I. *Revista Matemática Iberoamericana*, 32(1):275–376, 2016.
- [23] C. Fefferman, A. Israel, and G. K. Luli. Fitting a Sobolev function to data II. *Revista Matemática Iberoamericana*, 32(2):649–750, 2016.
- [24] C. Fefferman, A. Israel, and G. K. Luli. Fitting a Sobolev function to data III. *Revista Matemática Iberoamericana*, 32(3):1039–1126, 2016.
- [25] C. Fefferman and B. Klartag. Fitting a  $C^m$ -smooth function to data I. *Annals of Mathematics*, 169(1):315–346, 2009.
- [26] C. Fefferman and B. Klartag. Fitting a  $C^m$ -smooth function to data II. *Revista Matemática Iberoamericana*, 25(1):49–273, 2009.
- [27] I. Fonseca and G. Leoni. *Modern Methods in the Calculus of Variations:  $L^p$  Spaces*. Springer Science & Business Media, 2007.
- [28] N. García-Trillos. Variational limits of k-nn graph based functionals on data clouds. *arXiv preprint arXiv:1607.00696*, 2016.
- [29] N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev. Spectral convergence of the empirical graph Laplacian. *In Preparation*, 2017.
- [30] N. García Trillos and R. Murray. A new analytical approach to consistency and overfitting in regularized empirical risk minimization. *to appear in the European Journal of Applied Mathematics*, *arXiv preprint arXiv: 1607.00274*, 2016.
- [31] N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in  $\infty$ -transportation distance. *Canadian Journal of Mathematics*, 67:1358–1383, 2015.
- [32] N. García Trillos and D. Slepčev. Continuum limit of Total Variation on point clouds. *Archive for Rational Mechanics and Analysis*, 220(1):193–241, 2016.
- [33] N. García Trillos and D. Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 2016.
- [34] N. García Trillos, D. Slepčev, J. von Brecht, T. Laurent, and X. Bresson. Consistency of cheeger and ratio graph cuts. *Journal of Machine Learning Research*, 2015.

- [35] E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006.
- [36] M. Hein. Uniform convergence of adaptive graph-based regularization. In *International Conference on Computational Learning Theory*, pages 50–64, 2006.
- [37] M. Hein, J.-Y. Audibert, and U. von Luxburg. From graphs to manifolds—weak and strong pointwise consistency of graph Laplacians. In *Learning theory*, pages 470–485. Springer, 2005.
- [38] T. Leighton and P. Shor. Tight bounds for minimax grid matching with applications to the average case analysis of algorithms. *Combinatorica*, 9(2):161–187, 1989.
- [39] Z. Li and Z. Shi. A convergent point integral method for isotropic elliptic equations on a point cloud. *Multiscale Modeling & Simulation*, 14(2):874–905, 2016.
- [40] Z. Li, Z. Shi, and J. Sun. Point integral method for solving poisson-type equations on manifolds from point clouds with convergence guarantees. *Communications in Computational Physics*, 22(1):228–258, 2017.
- [41] B. Nadler, N. Srebro, and X. Zhou. Statistical analysis of semi-supervised learning: The limit of infinite unlabelled data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1330–1338, 2009.
- [42] B. Pelletier and P. Pudlo. Operator norm convergence of spectral clustering on level sets. *J. Mach. Learn. Res.*, 12:385–416, 2011.
- [43] M. Penrose. *Random Geometric Graphs*. Oxford University Press, 2003.
- [44] A. C. Ponce. A new approach to Sobolev spaces and connections to  $\Gamma$ -convergence. *Calc. Var. Partial Differential Equations*, 19(3):229–255, 2004.
- [45] F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87. Springer, 2015.
- [46] P. W. Shor and J. E. Yukich. Minimax grid matching and empirical measures. *Ann. Probab.*, 19(3):1338–1348, 1991.
- [47] A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.
- [48] A. Singer and H.-T. Wu. Spectral convergence of the connection Laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1):58–123, 2017.
- [49] Michel Talagrand. *Upper and lower bounds of stochastic processes*, volume 60 of *Modern Surveys in Mathematics*. Springer-Verlag, Berlin Heidelberg, 2014.
- [50] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev. A transportation  $L^p$  distance for signal analysis. *Journal of Mathematical Imaging and Vision*, 59(2):187–210, 2017.
- [51] M. Thorpe and D. Slepčev. Transportation  $L^p$  distances: Properties and extensions. *In preparation*, 2017.

- [52] M. Thorpe and F. Theil. Asymptotic analysis of the Ginzburg-Landau functional on point clouds. *to appear in the Proceedings of the Royal Society of Edinburgh Section A: Mathematics*, *arXiv preprint arXiv:1604.04930*, 2017.
- [53] M. Thorpe, F. Theil, A. M. Johansen, and N. Cade. Convergence of the  $k$ -means minimization problem using  $\Gamma$ -convergence. *SIAM Journal on Applied Mathematics*, 75(6):2444–2474, 2015.
- [54] D. Ting, L. Huang, and M. I. Jordan. An analysis of the convergence of graph Laplacians. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.
- [55] C. Villani. *Optimal Transport Old and New*. Springer-Verlag Berlin Heidelberg, 2009.
- [56] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 2007.
- [57] U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.
- [58] X. Wang. Spectral convergence rate of graph Laplacian. *arXiv preprint arXiv:1510.08110*, 2015.
- [59] D. Zhou and B. Schölkopf. Regularization on discrete spaces. In *Proceedings of the 27th DAGM Conference on Pattern Recognition*, PR’05, pages 361–368, Berlin, Heidelberg, 2005. Springer-Verlag.
- [60] X. Zhou and M. Belkin. Semi-supervised learning by higher order regularization. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 892–900, 2011.
- [61] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, 2003.