Accepted Manuscript

Large Data and Zero Noise Limits of Graph-Based Semi-Supervised Learning Algorithms

Matthew M. Dunlop, Dejan Slepčev, Andrew M. Stuart, Matthew Thorpe

PII: S1063-5203(18)30139-8
DOI: https://doi.org/10.1016/j.acha.2019.03.005
Reference: YACHA 1313

To appear in: *Applied and Computational Harmonic Analysis*

Received date: 25 May 2018
Revised date: 28 December 2018
Accepted date: 8 March 2019

Please cite this article in press as: M.M. Dunlop et al., Large Data and Zero Noise Limits of Graph-Based Semi-Supervised Learning Algorithms, *Appl. Comput. Harmon. Anal.* (2019), https://doi.org/10.1016/j.acha.2019.03.005

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Large Data and Zero Noise Limits of Graph-Based Semi-Supervised Learning Algorithms

Matthew M. Dunlop[a], Dejan Slepčev[b], Andrew M. Stuart[a], Matthew Thorpe[c]

[a] Computing and Mathematical Sciences, Caltech, Pasadena, CA 91125
[b] Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213
[c] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 0WA

Abstract

Scalings in which the graph Laplacian approaches a differential operator in the large graph limit are used to develop understanding of a number of algorithms for semi-supervised learning; in particular the extension, to this graph setting, of the probit algorithm, level set and kriging methods, are studied. Both optimization and Bayesian approaches are considered, based around a regularizing quadratic form found from an affine transformation of the Laplacian, raised to a, possibly fractional, exponent. Conditions on the parameters defining this quadratic form are identified under which well-defined limiting continuum analogues of the optimization and Bayesian semi-supervised learning problems may be found, thereby shedding light on the design of algorithms in the large graph setting. The large graph limits of the optimization formulations are tackled through $\Gamma-$convergence, using the recently introduced TL^p metric. The small labelling noise limits of the Bayesian formulations are also identified, and contrasted with pre-existing harmonic function approaches to the problem.

Keywords: Semi-supervised learning, Bayesian inference, higher-order fractional Laplacian, asymptotic consistency, kriging.
2010 MSC: 62G20, 62C10, 62F15, 49J55

ACCEPTED MANUSCRIPT

1. Introduction

1.1. Context

This paper is concerned with the semi-supervised learning problem of determining labels on an entire set of (feature) vectors $\{x_j\}_{j\in Z}$, given (possibly noisy) labels $\{y_j\}_{j\in Z'}$ on a subset of feature vectors with indices $j \in Z' \subset Z$. To be concrete we will assume that the x_j are elements of $\mathbb{R}^d$, $d \geq 2$, and consider the binary classification problem in which the y_j are elements of $\{\pm 1\}$. Our goal is to characterize algorithms for this problem in the large data limit where $n = |Z| \to \infty$; additionally we will study the limit where the noise in the label data disappears. Studying these limits yields insight into the classification problem and algorithms for it.

Semi-supervised learning as a subject has been developed primarily over the last two decades and the references [1, 2] provide an excellent source for the historical context. Graph based methods proceed by forming a graph with n nodes Z, and use the unlabeled data $\{x_j\}_{j\in Z}$ to provide an $n \times n$ weight matrix W quantifying the affinity of the nodes of the graph with one another. The labelling information on Z' is then spread to the whole of Z, exploiting these affinities. In the absence of labelling information we obtain the problem of unsupervised learning; for example the spectrum of the graph Laplacian L forms the basis of widely used spectral clustering methods [3, 4, 5]. Other approaches are combinatorial, and largely focussed on graph cut methods [6, 7, 8]. However relaxation and approximation are required to beat the combinatorial hardness of these problems [9] leading to a range of methods based on Markov random fields [10] and total variation relaxation [11]. In [2] a number of new approaches were introduced, including label propagation and the generalization of kriging, or Gaussian process regression [12], to the graph setting [13]. These regression methods opened up new approaches to the problem, but were limited in scope because the underlying real-valued Gaussian process was linked directly to the categorical label data which is (arguably) not natural from a modelling perspective; see [14] for a discussion of the distinctions between regression and

ACCEPTED MANUSCRIPT

classification. The logit and probit methods of classification [15] side-step this problem by postulating a link function which relates the underlying Gaussian process to the categorical data, amounting to a model linking the unlabeled and labeled data. The support vector machine [16] makes a similar link, but it lacks a natural probabilistic interpretation.

The probabilistic formulation is important when it is desirable to equip the classification with measures of uncertainty. Hence, we will concentrate on the probit algorithm in this paper, and variants on it, as it has a probabilistic formulation. The statement of the probit algorithm in the context of graph based semi-supervised learning may be found in [17]. An approach bridging the combinatorial and Gaussian process approaches is the use of Ginzburg-Landau models which work with real numbers but use a penalty to constrain to values close to the range of the label data $\{\pm 1\}$; these methods were introduced in [18], large data limits studied in [19, 20, 21], and given a probabilistic interpretation in [17]. Finally we mention the Bayesian level set method. This approach takes the idea of using level sets for inversion in the class of interface problems [22] and gives it a probabilistic formulation which has both theoretical foundations and leads to efficient algorithms [23]; classification may be viewed as an interface problem on a graph (a graph cut is an interface for example) and thus the Bayesian level set method is naturally extended to this setting as shown in [17]. As part of this paper we will show that the probit and Bayesian level set methods are closely related.

A significant challenge for the field, both in terms of algorithmic development, and in terms of fundamental theoretical understanding, is the setting in which the volume of unlabeled data is high, relative to the volume of labeled data. One way to understand this setting is through the study of large data limits in which $n = |Z| \to \infty$. This limit is studied in [24], and was addressed more recently under different assumptions in [25]. Both papers assume that the unlabeled data is drawn i.i.d. from a measure with Lebesgue density on a subset of $\mathbb{R}^d$, but the assumptions on graph construction differ: in [24] the graph bandwidth is fixed as $n \to \infty$ resulting in the limit of the graph Laplacian being

ACCEPTED MANUSCRIPT

a non-local operator, whilst in [25] the bandwidth vanishes in the limit resulting in the limit being a weighted Laplacian (divergence form elliptic operator).

In [26] it is demonstrated that algorithms based on use of the discrete Dirichlet energy computed from the graph Laplacian can behave poorly for $d \geq 2$, in the large data limit, if they attempt pointwise labelling. In [27] it is argued that use of quadratic forms based on powers $\alpha > \frac{d}{2}$ of the graph Laplacian can ameliorate this problem. Our work, which studies a range of algorithms all based on optimization or Bayesian formulations exploiting quadratic forms, will take this body of work considerably further, proving large data limit theorems for a variety of algorithms, and showing the role of the parameter α in this infinite data limit. In doing so we shed light on the difficult question of how to scale and tune algorithms for graph based semi-supervised learning; in particular we state limit theorems of various kinds which require, respectively, either $\alpha > \frac{d}{2}$ or $\alpha > d$ to hold. We also study the small noise limit and show how both the probit and Bayesian level set algorithms coincide and, furthermore, provide a natural generalization of the harmonic functions approach of [13, 28], a generalization which is arguably more natural from a modeling perspective.

Our large data limit theorems concern the maximum a posteriori (MAP) estimator rather than a Bayesian posterior distribution. However two remarkable recent papers [29, 30] demonstrate a methodology for proving limit theorems concerning Bayesian posterior distributions themselves, exploiting the variational characterization of Bayes theorem; extending the work in those papers to the algorithms considered in this paper would be of great interest.

1.2. Our Contribution

We derive a canonical continuum inverse problem which characterizes graph based semi-supervised learning: find function $u : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R}$ from knowledge of $\mathrm{sign}(u)$ on $\Omega' \subset \Omega$. [1] The latent variable u characterizes the unlabeled data

[1] We note that throughout the paper Ω is the physical domain, and not the set of events of a probability space.

ACCEPTED MANUSCRIPT

and its sign is the labeling information. This highly ill-posed inverse problem is potentially solvable because of the very strong prior information provided by the unlabeled data; we characterize this information via a mean zero Gaussian process prior on u with covariance operator $\mathcal{C} \propto (\mathcal{L} + \tau^2 I)^{-\alpha}$. The operator $\mathcal{L}$ is a weighted Laplacian found as a limit of the graph Laplacian, and as a consequence depends on the distribution of the unlabeled data.

In order to derive this canonical inverse problem we study the probit and Bayesian level set algorithms for semi-supervised learning. We build on the large unlabeled data limit setting of [25]. In this setting there is an intrinsic scaling parameter ε_n that characterizes the length scale on which edge weights between nodes are significant; the analysis identifies a lower bound on ε_n which is necessary in order for the graph to remain connected in the large data limit and under which the graph Laplacian L converges to a differential operator $\mathcal{L}$ of weighted Laplacian form. The work uses $\Gamma-$convergence in the TL^2 optimal transport metric, introduced in [25], and proves convergence of the quadratic form defined by L to one defined by $\mathcal{L}$. We make the following contributions which significantly extend this work to the semi-supervised learning setting.

- We prove $\Gamma-$convergence in TL^2 of the quadratic form defined by $(L + \tau^2 I)^\alpha$ to that defined by $(\mathcal{L} + \tau^2 I)^\alpha$ and identify parameter choices in which the limiting Gaussian measure with covariance $(\mathcal{L} + \tau^2 I)^{-\alpha}$ is well-defined. See Theorems 2.2, 2.5 and Proposition 2.6.

- We introduce large data limits of the probit and Bayesian level set problem formulations in which the volume of unlabeled data $n = |Z| \to \infty$, distinguishing between the cases where the volume of labeled data $|Z'|$ is fixed and where $|Z'|/n$ is fixed. See section 4 for the function space analogues of the graph based algorithms introduced in section 3.

- We use the theory of $\Gamma-$convergence to derive a continuum limit of the probit algorithm when employed in MAP estimation mode; this theory demonstrates the need for $\alpha > \frac{d}{2}$ and an upper bound on ε_n in the large

ACCEPTED MANUSCRIPT

data limit where the volume of labeled data $|Z'|$ is fixed. See Theorems 4.2 and 4.3

- We use the properties of Gaussian measures on function spaces to write down well defined limits of the probit and Bayesian level set algorithms, when employed in Bayesian probabilistic mode, to determine the posterior distribution on labels given observed data; this theory demonstrates the need for $\alpha > \frac{d}{2}$ in order for the limiting probability distribution to be meaningful for both large data limits; indeed, depending on the geometry of the domain from which the feature vectors are drawn, it may require $\alpha > d$ for the case where the volume of labeled data is fixed. See Theorem 2.5 and Proposition 2.6 for these conditions on α, and for details of the limiting probability measures see equations (21), (22), (23) and (24).

- We show that the probit and Bayesian level set methods have a common Bayesian inverse problem limit, mentioned above, by studying their weak limits as noise levels on the labeled data tends to zero. See Theorems 3.3 and 4.6.

- We provide numerical experiments which illusrate the large graph limits introduced and studied in this paper; see section 5.

1.3. Paper Structure

In section 2 we study a family of quadratic forms which arise naturally in all the algorithms that we study. By means of the $\Gamma-$convergence techniques pioneered in [25] we show that these quadratic forms have a limit defined by families of differential operators in which the finite graph parameters appear in an explicit and easily understood fashion. Section 3 is devoted to the definition of the three graph based algorithms that we study in this paper: the probit and Bayesian level set algorithms, and the graph analogue of kriging. In section 4 we write down the function space limits of these algorithms, obtained when the volume n of unlabeled data tends to infinity, and in the case of the maximum a posteriori estimator for probit use $\Gamma-$convergence to study large graph limits

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

rigorously; we also show that the probit and Bayesian level set algorithms have a common zero noise limit. Section 5 contains numerical experiments for the function space limits of the algorithms, in both optimization (MAP) and sampling (fully Bayesian MCMC) modalities. We conclude in section 6 with a summary and directions for future research. All proofs are given in the Appendix, section 7. This choice is made in order to separate the form and implications of the theory from the proofs; both the statements and proofs comprise the contributions of this work, but since they may be of interest to different readers they are separated, by use of the Appendix.

2. Key Quadratic Form and Its Limits

2.1. Graph Setting

From the unlabeled data $\{x_j\}_{j=1}^n$ we construct a weighted graph $G = (Z, W)$ where $Z = \{1, \cdots, n\}$ are the vertices of the graph and W the edge weight matrix; W is assumed to have entries $\{w_{ij}\}$ between nodes i and j given by

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

We will discuss the choice of the function $\eta_\varepsilon : \mathbb{R} \mapsto \mathbb{R}^+$ in detail below; heuristically it should be thought of as proportional to a mollified Dirac mass, or a characteristic function of a small interval. From W we construct the graph Laplacian as follows. We define the diagonal matrix $D = \text{diag}\{d_{ii}\}$ with entries $d_{ii} = \sum_{j \in Z} w_{ij}$. We can then define the unnormalized graph Laplacian $L = D - W$. Our results may be generalized to the normalized graph Laplacian $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ and we will comment on this in the conclusions.

2.2. Quadratic Form

We view $u : Z \mapsto \mathbb{R}$ as a vector in $\mathbb{R}^n$ and define the quadratic form

$$\langle u, Lu \rangle = \frac{1}{2} \sum_{i,j \in Z} w_{ij} |u(i) - u(j)|^2;$$

here $\langle \cdot, \cdot \rangle$ denotes the standard Euclidean inner-product on $\mathbb{R}^n$. This is the discrete Dirichlet energy defined via the graph Laplacian L which appears as a

ACCEPTED MANUSCRIPT

basic quantity in many unsupervised and semi-supervised learning algorithms. In this paper our interest focusses on forms based on powers of L:

$$J_n^{(\alpha,\tau)}(u) = \frac{1}{2n}\langle u, A^{(n)}u\rangle$$

where, for $\tau \geq 0$ and $\alpha > 0$,

$$A^{(n)} = (s_n L + \tau^2 I)^{\alpha}. \tag{1}$$

The sequence parameters s_n will be chosen appropriately to ensure that the quadratic form $J_n^{(\alpha,\tau)}(u)$ converges to a well-defined limit as $n \to \infty$.

In addition to working in a set-up which results in a well-defined limit, we will also ask that this limit results in a quadratic form defined by a differential operator. This, of course, requires some form of localization and we will encode this as follows: we will assume that $\eta_\varepsilon(\cdot) = \varepsilon^{-d}\eta(\cdot/\varepsilon)$, inducing a Dirac mass approximation as $\varepsilon \to 0$; later we will discuss how to relate ε to n. For now we state the assumptions on η that we employ throughout the paper:

Assumptions 1 (on η)**.** *The edge weight profile function η satisfies:*

(K1) $\eta(0) > 0$ and $\eta(\cdot)$ is continuous at 0;

(K2) η is non-increasing;

(K3) $\int_0^\infty \eta(r) r^{d+1} dr < \infty$;

Remark 2.1. *The prototypical example for η is $\eta(t) = 1$ if $|t| < 1$ and $\eta(t) = 0$ otherwise. In this example the graph has edges between any two nodes closer than ε; this is often referred to as the random geometric graph. Clearly this choice of η satisfies Assumptions 1.*

Notice that assumption (K3) implies that

$$\sigma_\eta := \frac{1}{d}\int_{\mathbb{R}^d} \eta(|h|)|h|^2 dh < \infty \quad \text{and} \quad \beta_\eta := \int_{\mathbb{R}^d} \eta(|h|) dh < \infty. \tag{2}$$

A notable fact about the limits that we study in the remainder of the paper is that they depend on η only through the constants σ_η, β_η, provided Assumptions 1 holds and $\varepsilon = \varepsilon_n$ and s_n are chosen as appropriate functions of n.

ACCEPTED MANUSCRIPT

2.3. Limiting Quadratic Form

The limiting quadratic form is defined on an open and bounded set $\Omega \subset \mathbb{R}^d$.

Assumptions 2 (on Ω)**.** *We assume that Ω is a connected, open and bounded subset of $\mathbb{R}^d$. We also assume that Ω has $C^{1,1}$boundary.* [2]

Assumptions 3 (on density ρ)**.** *We assume that n feature vectors $x_j \in \Omega$ are sampled i.i.d. from a probability measure μ supported on Ω with smooth Lebesgue density ρ bounded above and below by finite strictly positive constants $\rho^{\pm}$ uniformly on $\overline{\Omega}$.*

We index the data by $Z = \{1, \cdots, n\}$ and let $\Omega_n = \{x_i\}_{i \in Z}$ be the data set. This data set induces the empirical measure

$$\mu_n = \frac{1}{n} \sum_{i \in Z} \delta_{x_i}.$$

Given a measure ν on Ω we define the weighted Hilbert space $L^2_\nu = L^2_\nu(\Omega; \mathbb{R})$ with inner-product

$$\langle a, b \rangle_\nu = \int_\Omega a(x) b(x) \nu(dx) \tag{3}$$

and the induced norm defined by the identity $\| \cdot \|^2_{L^2_\nu} = \langle \cdot, \cdot \rangle_\nu$. Note that with these definitions we have

$$J_n^{(\alpha,\tau)} : L^2_{\mu_n} \mapsto [0, +\infty), \qquad J_n^{(\alpha,\tau)}(u) = \frac{1}{2} \langle u, A^{(n)} u \rangle_{\mu_n}.$$

In what follows we apply a form of $\Gamma-$convergence to establish that for large n the quadratic form $J_n^{(\alpha,\tau)}$ is well approximated by the limiting quadratic form

$$J_\infty^{(\alpha,\tau)} : L^2_\mu \mapsto [0, +\infty) \cup \{+\infty\}, \qquad J_\infty^{(\alpha,\tau)}(u) = \frac{1}{2} \langle u, \mathcal{A} u \rangle_\mu.$$

---

[2] The assumption that Ω is connected is not essential but makes stating the results simpler. We remark that a number of the results, and in particular the convergence of Theorem 2.2, hold if we only assume that the boundary of Ω is Lipschitz. We need the stronger assumption in order to be able to employ elliptic regularity to characterize functions in fractional Sobolev spaces, see Section 2.4 and Lemma 7.1; this is essential to be able to define Gaussian measures on function spaces, and therefore needed to define a Bayesian approach in which uncertainty of classifiers may be estimated.

ACCEPTED MANUSCRIPT

Here μ is the measure on Ω with density ρ, and we define the L^2_μ self-adjoint differential operator $\mathcal{L}$ by

$$\mathcal{L}u = -\frac{1}{\rho}\nabla\cdot(\rho^2\nabla u), \quad x\in\Omega, \qquad \frac{\partial u}{\partial n} = 0, \quad x\in\partial\Omega. \tag{4}$$

The operator $\mathcal{A}$ is then defined by $\mathcal{A} = (\mathcal{L} + \tau^2 I)^\alpha$.

We may now relate the quadratic forms defined by $A^{(n)}$ and $\mathcal{A}$. The TL^2 topology is introduced in [25] and defined in the Appendix section 7.2.2 for convenience. The following theorem is proved in section 7.4.

Theorem 2.2. *Let Assumptions 1–3 hold. Let $\alpha > 0$, $\{\varepsilon_n\}_{n=1,2,\dots}$ be a positive sequence converging to zero, and such that*

$$\begin{aligned} \lim_{n\to\infty}\left(\frac{\log n}{n}\right)^{1/d}\frac{1}{\varepsilon_n} &= 0 \qquad \text{if } d\geq 3,\\ \lim_{n\to\infty}\left(\frac{\log n}{n}\right)^{1/2}\frac{(\log n)^{\frac{1}{4}}}{\varepsilon_n} &= 0 \qquad \text{if } d = 2, \end{aligned} \tag{5}$$

and assume that the scale factor s_n is defined by

$$s_n = \frac{2}{\sigma_\eta n\varepsilon_n^2}. \tag{6}$$

Then, with probability one, we have

1. *$\Gamma\text{-}\lim_{n\to\infty} J_n^{(\alpha,\tau)} = J_\infty^{(\alpha,\tau)}$ with respect to the TL^2 topology;*
2. *if $\tau = 0$, any sequence $\{u_n\}$ with $u_n : \Omega_n \to \mathbb{R}$ satisfying $\sup_n \|u_n\|_{L^2_{\mu_n}} < \infty$ and $\sup_{n\in\mathbb{N}} J_n^{(\alpha,0)}(u_n) < \infty$ is pre-compact in the TL^2 topology;*
3. *if $\tau > 0$, any sequence $\{u_n\}$ with $u_n : \Omega_n \to \mathbb{R}$ satisfying $\sup_{n\in\mathbb{N}} J_n^{(\alpha,\tau)}(u_n) < \infty$ is pre-compact in the TL^2 topology.*

Remark 2.3. *As we discuss in section 7.2.1 of the appendix, Γ-convergence and pre-compactness allow one to show that minimizers of a sequence of functionals converge to the minimizer of the limiting functional. The results of Theorem 2.2 provide the Γ-convergence and pre-compactness of fractional Dirichlet energies, which are the key term of the functionals, such as* (10) *below, that define the learning algorithms that we study. In particular Theorem 2.2 enables us to prove the convergence, in the large data limit $n\to\infty$, of minimizers of functionals such as* (10) *(i.e. of outcomes of learning algorithms), as shown in Theorem 4.2.*

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

2.4. Function Spaces

The operator $\mathcal{L}$ given by (4) is uniformly elliptic as a consequence of the assumptions on ρ, and is self-adjoint with respect to the inner product (3) on L^2_μ. By standard theory, it has a discrete spectrum: $0 = \lambda_1 < \lambda_2 \leq \cdots$, where the fact that $0 < \lambda_2$ uses the connectedness of the domain and the uniform positivity of ρ on the domain. Let φ_i for $i = 1, \ldots$ be the associated L^2_μ-orthonormal eigenfunctions. They form a basis of L^2_μ.

By Weyl's law the eigenvalues of $\{\lambda_j\}_{j\geq 1}$ of $\mathcal{L}$ satisfy $\lambda_j \asymp j^{2/d}$. For completeness a simple proof is proved in Lemma 7.10; the analogous and more general results applicable to the Laplace-Beltrami operator may be found in, Hörmander [31].

Spectrally defined Sobolev spaces. For $s \geq 0$ we define

$$\mathcal{H}^s(\Omega) = \left\{ u \in L^2_\mu \ : \ \sum_{k=1}^{\infty} \lambda_k^s a_k^2 < \infty \right\},$$

where $a_k = \langle u, \varphi_k \rangle_\mu$ and thus $u = \sum_k a_k \varphi_k$ in L^2_μ. We note that $\mathcal{H}^s(\Omega)$ is a Hilbert space with respect to the inner product

$$\langle\!\langle u, v \rangle\!\rangle_{s,\mu} = a_1 b_1 + \sum_{k=2}^{\infty} \lambda_k^s a_k b_k$$

where $b_k = \langle v, \varphi_k \rangle_\mu$. It follows from the definition that for any $s \geq 0$, $\mathcal{H}^s(\Omega)$ is isomorphic to a weighted $\ell^2(\mathbb{N})$ space, where the weights are formed by the sequence $1, \lambda_2^s, \lambda_3^s, \ldots$.

In Lemma 7.1 in the Appendix section 7.1 we show that for any integer $s > 0$, $\mathcal{H}^s(\Omega) \subset H^s(\Omega)$ where $H^s(\Omega)$ is the standard fractional Sobolev space. More precisely we characterize $\mathcal{H}^s(\Omega)$ as the set of those functions in $H^s(\Omega)$ which satisfy the appropriate boundary condition and show that the norms of $\mathcal{H}^s(\Omega)$ and $H^s(\Omega)$ are equivalent on $\mathcal{H}^s(\Omega)$.

We also note that for any integer s and $\theta \in (0,1)$ the space $\mathcal{H}^{s+\theta}$ is a interpolation space between $\mathcal{H}^s$ and $\mathcal{H}^{s+1}$. In particular $\mathcal{H}^{s+\theta} = [\mathcal{H}^s, \mathcal{H}^{s+1}]_{\theta,2}$, where the real interpolation space used is as in Definition 3.3 of Abels [32]. This identification of $\mathcal{H}^s$ follows from the characterization of interpolation spaces of

ACCEPTED MANUSCRIPT

weighted L^p spaces by Peetre [33], as referenced by Gilbert [34]. Together these facts allow us to characterize the Hölder regularity of functions in $\mathcal{H}^s(\Omega)$.

Lemma 2.4. *Under Assumptions 2–3, for all $s \geq 0$ there exists a bounded, linear, extension mapping $E : \mathcal{H}^s(\Omega) \to H^s(\mathbb{R}^d)$. That is for all $f \in \mathcal{H}^s(\Omega)$, $E(f)|_\Omega = f$ a.e. Furthermore:*

(i) if $s < \frac{d}{2}$ then $\mathcal{H}^s(\Omega)$ embeds continuously in $L^q(\Omega)$ for any $q \leq \frac{2d}{d-2s}$;

(ii) if $s > \frac{d}{2}$ then $\mathcal{H}^s(\Omega)$ embeds continuously in $C^{0,\gamma}(\Omega)$ for any $\gamma < \min\{1, s - \frac{d}{2}\}$.

The proof is presented in the Appendix 7.1.

We note that this implies that when $\alpha > \frac{d}{2}$ pointwise evaluation is well-defined in the limiting quadratic form $J_\infty^{(\alpha,\tau)}$; this will be used in what follows to show that the the limiting labelling model obtained when $|Z'|$ is fixed is well-posed.

2.5. Gaussian Measures of Function Spaces

Using the ellipticity of $\mathcal{L}$, Weyl's law, and Lemma 2.4 allows us to characterize the regularity of samples of Gaussian measures on L^2_μ. The proof of the following theorem is a straightforward application of the techniques in [35, Theorem 2.10] to obtain the Gaussian measures on $\mathcal{H}^s(\Omega)$. Concentration of the measure on H^s and on $C^{0,\gamma}(\Omega)$ then follows from Lemma 2.4. When $\tau = 0$ we work on the space orthogonal to constants in order that $\mathcal{C}$ (defined in the theorem below) is well defined.

Theorem 2.5. *Let Assumptions 2–3 hold. Let $\mathcal{L}$ be the operator defined in (4), and define $\mathcal{C} = (\mathcal{L} + \tau^2 I)^{-\alpha}$. For any fixed $\alpha > \frac{d}{2}$ and $\tau \geq 0$, the Gaussian measure $N(0, \mathcal{C})$ is well-defined on L^2_μ. Draws from this measure are almost surely in $H^s(\Omega)$ for any $s < \alpha - \frac{d}{2}$, and consequently in $C^{0,\gamma}(\Omega)$ for any $\gamma < \min\{1, \alpha - d\}$ if $\alpha > d$.*

We note that if the operator $\mathcal{L}$ has eigenvectors which are as regular as those of the Laplacian on a flat torus then the conclusions of Theorem 2.5 can

ACCEPTED MANUSCRIPT

be strengthened. Namely if in addition to what we know about $\mathcal{L}$, there is $C > 0$ such that

$$\sup_{j\geq 1}\left(\|\varphi_j\|_{L^\infty} + \frac{1}{j^{1/d}}\mathrm{Lip}(\varphi_j)\right) \leq C, \tag{7}$$

then the Kolmogorov continuity technique [35, Section 7.2.5] can be used to show additional Hölder continuity.

Proposition 2.6. *Let Assumptions 2–3 hold. Assume the operator $\mathcal{L}$ satisfies condition (7) and define $\mathcal{C} = (\mathcal{L} + \tau^2 I)^{-\alpha}$. For any fixed $\alpha > d/2$ and $\tau \geq 0$, the Gaussian measure $N(0, \mathcal{C})$ is well-defined on L^2_μ. Draws from this measure are almost surely in $H^s(\Omega;\mathbb{R})$ for any $s < \alpha - d/2$, and in $C^{0,\gamma}(\Omega;\mathbb{R})$ for any $\gamma < \min\{1, \alpha - \frac{d}{2}\}$ if $\alpha > \frac{d}{2}$.*

We note that in general one cannot expect that the operator $\mathcal{L}$ satisfies the bound (7). For example, for the ball there is a sequence of eigenfunctions which satisfy $\|\varphi_k\|_{L^\infty} \sim \lambda_k^{(d-1)/4} \sim k^{(d-2)/(2d)}$, see [36]. In fact this is the largest growth of eigenfunctions possible, as on general domains with smooth boundary $\|\varphi_k\|_{L^\infty} \lesssim \lambda_k^{(d-1)/4}$, as follows from the work of Grieser, [36]. Analogous bounds have first been established for operators on manifolds without boundary by Hörmander, [31]. This bound is rarely saturated as shown by Sogge and Zeldtich [37], but determining the scaling for most sets and manifolds remains open. Establishing the conditions on Ω under which the Theorem 2.5 can be strengthened as in Proposition 2.6 is of great interest.

3. Graph Based Formulations

We now assume that we have access to label data defined as follows. Let $\Omega' \subset \Omega$ and let $\Omega^\pm$ be two subsets of Ω' such that

$$\Omega^+ \cup \Omega^- = \Omega', \quad \overline{\Omega^+} \cap \overline{\Omega^-} = \varnothing.$$

We will consider two labelling scenarios:

- **Labelling Model 1**. $|Z'|/n \to \mathfrak{r} \in (0, \infty)$. We assume that $\Omega^\pm$ have positive Lebesgue measure. We assume that the $\{x_j\}_{j\in\mathbb{N}}$ are drawn i.i.d.

ACCEPTED MANUSCRIPT

from measure μ. Then if $x_j \in \Omega^+$ we set $y_j = 1$ and if $x_j \in \Omega^-$ then $y_j = -1$. The label variables y_j are not defined if $x_j \in \Omega \backslash \Omega'$ where $\Omega' = \Omega^+ \cup \Omega^-$. We assume $\mathrm{dist}(\Omega^+, \Omega^-) > 0$ and define $Z' \subset Z$ to be the subset of indices for which we have labels.

Labelling Model 2. $|Z'|$ fixed as $n \to \infty$. We assume that $\Omega^\pm$ comprise a fixed number of points, $n^\pm$ respectively. We assume that the $\{x_j\}_{j > n^+ + n^-}$ are drawn i.i.d. from measure μ whilst $\{x_j\}_{1 \le j \le n^+}$ are a fixed set of points in Ω^+ and $\{x_j\}_{n^+ + 1 \le j \le n^+ + n^-}$ are a fixed set of points in Ω^-. We label these fixed points by $y : \Omega^\pm \mapsto \{\pm 1\}$ as in **Labelling Model 1**. We define $Z' \subset Z$ to be the subset of indices $\{1, \cdots, n^+ + n^-\}$ for which we have labels and $\Omega' = \Omega^+ \cup \Omega^-$.

In both cases $j \in Z'$ if and only if $x_j \in \Omega'$. But in Model 1 the x_j are drawn i.i.d. and assigned labels when they lie in Ω', assumed to have positive Lebesgue measure; in Model 2 the $\{(x_j, y_j)\}_{j \in Z'}$ are provided, in a possibly non-random way, independently of the unlabeled data.

We will identify $u \in \mathbb{R}^n$ and $u \in L^2_{\mu_n}(\Omega; \mathbb{R})$ by $u_j = u(x_j)$ for each $j \in Z$. Similarly, we will identify $y \in \mathbb{R}^{n^+ + n^-}$ and $y \in L^2_{\mu_n}(\Omega'; \mathbb{R})$ by $y_j = y(x_j)$ for each $j \in Z'$. We may therefore write, for example,

$$\frac{1}{n} \langle u, Lu \rangle_{\mathbb{R}^n} = \langle u, Lu \rangle_{\mu_n}$$

where u is viewed as a vector on the left-hand side and a function on Z on the right-hand side.

The algorithms that we study in this paper have interpretations through both optimization and probability. The labels are found from a real-valued function $u : Z \mapsto \mathbb{R}$ by setting $y = S \circ u : Z \mapsto \mathbb{R}$ with S the sign function defined by

$$S(0) = 0; \quad S(u) = 1,\ u > 0; \quad \text{and} \quad S(u) = -1,\ u < 0.$$

The objective function of interest takes the form

$$\mathsf{J}^{(n)}(u) = \frac{1}{2} \langle u, A^{(n)} u \rangle_{\mu_n} + r_n \Phi^{(n)}(u).$$

ACCEPTED MANUSCRIPT

The quadratic form depends only on the unlabeled data, while the function $\Phi^{(n)}$ is determined by the labeled data. Choosing $r_n = \frac{1}{n}$ in **Labeling Model 1** and $r_n = 1$ in **Labeling Model 2** ensures that the total labelling information remains of $\mathcal{O}(1)$ in the large n limit. Probability distributions constructed by exponentiating multiples of $\mathsf{J}^{(n)}(u)$ will be of interest to us; the probability is then high where the objective function is small, and vice-versa. Such probabilities represent the Bayesian posterior distribution on the conditional random variable $u|y$.

3.1. Probit

The probit algorithm on a graph is defined in [17] and here generalized to a quadratic form based on $A^{(n)}$ rather than L. We define

$$\Psi(v;\gamma) = \frac{1}{\sqrt{2\pi\gamma^2}}\int_{-\infty}^{v}\exp\left(-t^2/2\gamma^2\right)\mathrm{d}t \tag{8}$$

and then

$$\Phi_{\mathrm{p}}^{(n)}(u;\gamma) = -\sum_{j\in Z'}\log\Big(\Psi(y_j u_j;\gamma)\Big). \tag{9}$$

The function Ψ and its logarithm are shown in Figure 1 in the case $\gamma = 1$. The probit objective function is

$$\mathsf{J}_{\mathrm{p}}^{(n)}(u) = J_n^{(\alpha,\tau)}(u) + r_n\Phi_{\mathrm{p}}^{(n)}(u;\gamma), \tag{10}$$

where $r_n = \frac{1}{n}$ in **Labeling Model 1** and $r_n = 1$ in **Labeling Model 2**. The proof of Proposition 1 in [17] is readily modified to prove the following.

Proposition 3.1. *Let $\alpha > 0$, $\tau \geq 0$, $\gamma > 0$ and $r_n > 0$. Then $\mathsf{J}_{\mathrm{p}}^{(n)}$, defined by* (8-10)*, is strictly convex.*

It is also straightforward to check, by expanding u in the basis given by eigenvectors of $A^{(n)}$, that $\mathsf{J}_{\mathrm{p}}^{(n)}$ is coercive. This is proved by establishing that $J_n^{(\alpha,\tau)}$ is coercive on the orthogonal complement of the constant function. The coercivity in the remaining direction is provided by $\Phi_{\mathrm{p}}^{(n)}(u;\gamma)$ using the fact that Ω^+ and Ω^- are nonempty. Consequently $\mathsf{J}_{\mathrm{p}}^{(n)}$ has a unique minimizer;

ACCEPTED MANUSCRIPT

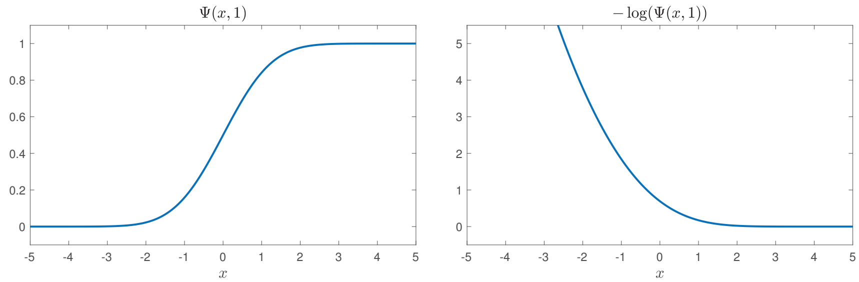

Figure 1: The function $\Psi(\cdot; 1)$, defined by (8), and its logarithm, which appears in the probit objective function.

Lemma 4.1 has the proof of the continuum analog of this; the proof on a graph is easily reconstructed from this.

The probabilistic analogue of the optimization problem for $\mathsf{J}_{\mathrm{p}}^{(n)}$ is as follows. We let $\nu_0^{(n)}(\mathrm{d}u; r)$ denote the centred Gaussian with covariance $C = r_n(A^{(n)})^{-1}$ (with respect to the inner product $\langle\cdot,\cdot\rangle_{\mu_n}$). We assume that the latent variable u is a priori distributed according to measure $\nu_0^{(n)}(\mathrm{d}u; r_n)$. If we then define the likelihood $y|u$ through the generative model

$$y_j = S\big(u_j + \xi_j\big) \tag{11}$$

with $\xi_j \overset{\text{iid}}{\sim} N(0, \gamma^2)$ then the posterior probability on $u|y$ is given by

$$\nu_{\mathrm{p}}^{(n)}(\mathrm{d}u) = \frac{1}{Z_{\mathrm{p}}^{(n)}} e^{-\Phi_{\mathrm{p}}^{(n)}(u;y)} \nu_0^{(n)}(\mathrm{d}u; r_n) \tag{12}$$

with $Z_{\mathrm{p}}^{(n)}$ the normalization to a probability measure. The measure $\nu_{\mathrm{p}}^{(n)}$ has Lebesgue density proportional to $e^{-r_n^{-1}\mathsf{J}_{\mathrm{p}}^{(n)}(u)}$.

3.2. *Bayesian Level Set*

We now define

$$\Phi_{\mathrm{ls}}^{(n)}(u; \gamma) = \frac{1}{2\gamma^2} \sum_{j\in Z'} \big|y_j - S\big(u_j\big)\big|^2. \tag{13}$$

The relevant objective function is

$$\mathsf{J}_{\mathrm{ls}}^{(n)}(u) = J_n^{(\alpha,\tau)}(u) + r_n \Phi_{\mathrm{ls}}^{(n)}(u; \gamma),$$

where again $r_n = \frac{1}{n}$ in **Labeling Model 1** and $r_n = 1$ in **Labeling Model 2**. We have the following:

ACCEPTED MANUSCRIPT

Proposition 3.2. *The infimum of of* $\mathsf{J}_{\mathrm{ls}}^{(n)}$ *is not attained.*

This follows using the argument introduced in a related context in [23]: assuming that a non-zero minimizer does exist leads to a contradiction upon multiplication of that minimizer by any number less than one; and zero does not achieve the infimum.

We modify the generative model (11) slightly to read

$$y_j = S\big(u_j\big) + \xi_j,$$

where now $\xi_j \overset{\mathrm{iid}}{\sim} N(0, r_n^{-1}\gamma^2)$. In this case, because the noise is additive, multiplying the objective function by r_n simply results in a rescaling of the observational noise; multiplication by r_n does not have such a simple interpretation in the case of probit. As a consequence the resulting Bayesian posterior distribution has significant differences with the probit case: the latent variable u is now assumed a priori to be distributed according to measure $\nu_0^{(n)}(\mathrm{d}u;1)$ Then

$$\nu_{\mathrm{ls}}^{(n)}(\mathrm{d}u) = \frac{1}{Z_{\mathrm{ls}}^{(n)}} e^{-r_n \Phi_{\mathrm{ls}}^{(n)}(u;\gamma)} \nu_0^{(n)}(\mathrm{d}u;1) \tag{14}$$

where $\nu_0^{(n)}$ is the same centred Gaussian as in the probit case. Note that $\nu_{\mathrm{ls}}^{(n)}$ is also the measure with Lebesgue density proportional to $e^{-\mathsf{J}_{\mathrm{ls}}^{(n)}(u)}$.

3.3. Small Noise Limit

When the size of the noise on the labels is small, the probit and Bayesian level set approaches behave similarly. More precisely, the measures $\nu_{\mathrm{p}}^{(n)}$ and $\nu_{\mathrm{ls}}^{(n)}$ share a common weak limit as $\gamma \to 0$. The following result is given without proof – this is because its proof is almost identical to that arising in the continuum limit setting of Theorem 4.6(ii) given in the appendix; indeed it is technically easier due to the fully discrete setting. Here $\Rightarrow$ denotes the weak convergence of probability measures.

Theorem 3.3. *Let* $\nu_0^{(n)}(\mathrm{d}u)$ *denote a Gaussian measure of the form* $\nu_0^{(n)}(\mathrm{d}u;r)$ *for any* r*, possibly depending on* n*. Define the set*

$$B_n = \{u \in \mathbb{R}^n \,|\, y_j u_j > 0 \,\text{for each}\, j \in Z'\}$$

ACCEPTED MANUSCRIPT

and the probability measure

$$\nu^{(n)}(\mathrm{d}u) = \mathsf{Z}^{-1}\mathbb{1}_{B_n}(u)\nu_0^{(n)}(\mathrm{d}u)$$

where $\mathsf{Z} = \nu_0^{(n)}(B_n)$*. Consider the posterior measures* $\nu_{\mathrm{p}}^{(n)}$ *defined in* (12) *and* $\nu_{\mathrm{ls}}^{(n)}$ *defined in* (14)*. Then* $\nu_{\mathrm{p}}^{(n)} \Rightarrow \nu^{(n)}$ *and* $\nu_{\mathrm{ls}}^{(n)} \Rightarrow \nu^{(n)}$ *as* $\gamma \to 0$*.*

3.4. Kriging

Instead of classification, where the sign of the latent variable u is made to agree with the labels, one can alternatively consider regression where u itself is made to agree with the labels [13, 28]. We consider this situation numerically in section 5. Here the objective is to

$$\text{minimize } \mathsf{J}_{\mathrm{k}}^{(n)}(u) := J_n^{(\alpha,\tau)}(u) \text{ subject to } u(x_j) = y_j \text{ for all } j \in Z'.$$

In the continuum setting this minimization is referred to as kriging, and we extend the terminology to our graph based setting. Kriging may also be defined in the case where the constraint is enforced as a soft least squares penalty; however we do not discuss this here.

The probabilistic analogue of this problem can be linked with the original work of Zhu et al [13, 28] which based classification on a centred Gaussian measure with inverse covariance given by the graph Laplacian, conditioned to take the value exactly 1 on labeled nodes where $y_j = 1$, and to take the value exactly -1 on labeled nodes where $y_j = -1$.

4. Function Space Limits of Graph Based Formulations

In this section we state $\Gamma-$limit theorems for the objective functions appearing in the probit algorithm. The proofs are given in the appendix. They rely on arguments which use the fact that we study perturbations of the $\Gamma-$limit theorem for the quadratic forms stated in section 2. We also write down formal infinite dimensional formulations of the probit and Bayesian level set posterior distributions, although we do not prove that these limits are attained. We do, however, show that the probit and level set posteriors have a common limit as $\gamma \to 0$, as they do on a finite graph.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

4.1. Probit

Under **Labelling Model 1**, the natural continuum limit of the probit objective functional is

$$\mathsf{J}_{\mathrm{p}}(v) = J_\infty^{(\alpha,\tau)}(v) + \Phi_{\mathrm{p},1}(v;\gamma) \tag{15}$$

where

$$\Phi_{\mathrm{p},1}(v;\gamma) = -\int_{\Omega'} \log(\Psi(y(x)v(x);\gamma))\,\mathrm{d}\mu(x) \tag{16}$$

for a given measurable function $y:\Omega' \to \{\pm 1\}$. For any $v \in L^2_\mu$, $\log(\Psi(y(x)v(x);\gamma))$ is integrable by Corollary 7.9. The proof of the following theorem is given in the appendix, in section 7.5.

Lemma 4.1. *Let Assumptions 1–3 hold. For $\alpha \geq 1$ and $\tau \geq 0$, consider the functional J_{p} with* **Labelling Model 1** *defined by* (15)*. Then, the functional J_{p} has a unique minimizer in $\mathcal{H}^\alpha(\Omega)$.*

Proof. Convexity of J_{p} follows from the proof of Proposition 1 in [17]. Let $\bar{v}_+$ and $\bar{v}_-$ be the averages of v on Ω_+ and Ω_- respectively. Namely let $\bar{v}_\pm = \frac{1}{|\Omega_\pm|}\int_{\Omega_\pm} v(x)\,\mathrm{d}x$. Note that

$$\mathsf{J}_{\mathrm{p}}(v) \geq J_\infty^{(\alpha,\tau)}(v) \geq \lambda_2^{\alpha-1} J_\infty^{(1,0)}(v) = -\frac{1}{2}\lambda_2^{\alpha-1}\int_\Omega v\nabla\cdot(\rho^2\nabla v)\,\mathrm{d}x \geq \frac{(\rho^-)^2\lambda_2^{\alpha-1}}{2}\|\nabla v\|^2_{L^2(\Omega)}.$$

Using the form of Poincaré inequality given in Theorem 13.27 of [38] implies that

$$\mathsf{J}_{\mathrm{p}}(v) \gtrsim \|\nabla v\|^2_{L^2(\Omega)} \gtrsim \int_\Omega |v-\bar{v}_+|^2 + |v-\bar{v}_-|^2\,\mathrm{d}x. \tag{17}$$

The convexity of $\Phi_{\mathrm{p},1}(v;\gamma)$ implies that

$$\Phi_{\mathrm{p},1}(v;\gamma) \geq -\log(\Psi(\bar{v}_+);\gamma)\mu(\Omega_+) - \log(\Psi(-\bar{v}_-);\gamma)\mu(\Omega_-)$$

Using that $\lim_{s\to-\infty} -\log(\Psi(s;\gamma)) = \infty$ we see that a bound on $\Phi_{\mathrm{p},1}(v;\gamma)$ provides a lower bound on $\bar{v}_+$ and an upper bound on $\bar{v}_-$. To see this let Θ be the inverse of $s \mapsto -\log(\Psi(s;\gamma))$. The preceding shows that

$$\bar{v}_+ \geq \Theta\left(\frac{\Phi_{\mathrm{p},1}(v;\gamma)}{\mu(\Omega_+)}\right) \geq \Theta\left(\frac{\mathsf{J}_{\mathrm{p}}(v)}{\mu(\Omega_+)}\right) \quad \text{and} \quad \bar{v}_- \leq -\Theta\left(\frac{\Phi_{\mathrm{p},1}(v;\gamma)}{\mu(\Omega_-)}\right) \leq -\Theta\left(\frac{\mathsf{J}_{\mathrm{p}}(v)}{\mu(\Omega_-)}\right).$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Let $c = \max\left\{-\Theta\left(\frac{\mathsf{J}_{\mathrm{p}}(v)}{\mu(\Omega_+)}\right), -\Theta\left(\frac{\mathsf{J}_{\mathrm{p}}(v)}{\mu(\Omega_-)}\right), 0\right\}$. Then $\bar{v}_+ \geq -c$ and $\bar{v}_- \leq c$. Using that, for any $a \in \mathbb{R}$, $v^2 \leq 2|v-a|^2 + 2a^2$, we obtain

$$\begin{aligned}
\int_\Omega v^2(x)\,\mathrm{d}x &\leq \int_{\{v(x)\leq -c\}} v^2(x)\,\mathrm{d}x + \int_{\{v(x)\geq c\}} v^2(x)\,\mathrm{d}x + c^2|\Omega| \\
&\leq 2\int_{\{v(x)\leq -c\}} |v+c|^2 + c^2\,\mathrm{d}x + 2\int_{\{v(x)\geq c\}} |v-c|^2 + c^2\,\mathrm{d}x + c^2|\Omega| \\
&\leq 5c^2|\Omega| + 2\int_{\{v(x)\leq -c\}} |v-\bar{v}_+|^2\,\mathrm{d}x + 2\int_{\{v(x)\geq c\}} |v-\bar{v}_-|^2\,\mathrm{d}x \\
&\lesssim c^2|\Omega| + \mathsf{J}_{\mathrm{p}}(v).
\end{aligned}$$

Then $\|v\|_{L^2}$ is bounded by a function of $\mathsf{J}_{\mathrm{p}}(v)$ and Ω.

Combining with (17) implies that a function of $\mathsf{J}_{\mathrm{p}}(v)$ bounds $\|v\|^2_{\mathcal{H}^\alpha(\Omega)}$ which establishes the coercivity of J_{p}. The functional J_{p} is weakly lower-semicontinuous in $\mathcal{H}^\alpha$, due to the convexity of both $J_\infty^{(\alpha,\tau)}$ and $\Phi_{\mathrm{p},1}$. Thus the direct method of the calculus of variations proves that J_{p} has a unique minimizer in $\mathcal{H}^\alpha(\Omega)$. □

The following theorem is proved in section 7.5.

Theorem 4.2. *Let the assumptions of* **Labelling Model 1** *and Theorem 2.2 hold with $\tau \geq 0$. Then, with probability one, any sequence of minimizers v_n of $\mathsf{J}_{\mathrm{p}}^{(n)}$ converge in TL^2 to v_∞, the unique minimizer of J_{p} in L^2_μ, and furthermore $\lim_{n\to\infty} \mathsf{J}_{\mathrm{p}}^{(n)}(v_n) = \mathsf{J}_{\mathrm{p}}(v_\infty) = \min_{v\in L^2_\mu} \mathsf{J}_{\mathrm{p}}(v)$.*

The analogous result under **Labelling Model 2**, i.e. convergence of minimizers, is an open question. In this case the natural continuum limit of the probit objective functional is

$$\mathsf{J}_{\mathrm{p}}(v) = J_\infty^{(\alpha,\tau)}(v) + \Phi_{\mathrm{p},2}(v;\gamma) \tag{18}$$

where

$$\Phi_{\mathrm{p},2}(v;\gamma) = -\sum_{j\in Z'} \log(\Psi(y(x_j)u(x_j);\gamma)) \tag{19}$$

for a given measurable function $y : \Omega' \to \{\pm 1\}$. When $\alpha \leq \frac{d}{2}$ this limiting model is not well-posed. In particular the regularity of the functional is not sufficient to impose pointwise data. More precisely, when $\alpha \leq \frac{d}{2}$ then there

ACCEPTED MANUSCRIPT

exists a sequence of smooth functions $v_k \in C^\infty(\Omega)$ such that $\lim_{k\to\infty} \mathsf{J}_\mathrm{p}(v_k) = 0$. In particular when $\alpha < \frac{d}{2}$, consider a smooth, compactly supported, mollifier ζ, with $\zeta(0) > 0$ and define $v_k(x) = c_k \sum_{i=1}^N y(x_i)\zeta_{1/k}(x - x_i)$ where $c_k \to \infty$ sufficiently slowly. Then $\Phi_{\mathrm{p},2}(v_k;\gamma) \to 0$ as $k \to \infty$ and, by a simple scaling argument (for appropriate c_k), $J_\infty^{(\alpha,\tau)}(v_k) \to 0$ as $k \to \infty$. Another way to see that the problem is not well defined is that the functions in $\mathcal{H}^\alpha(\Omega)$ (which is the natural space to consider J_p on) are not continuous in general and evaluating $\Phi_{\mathrm{p},2}(v;\gamma)$ is not well defined.

When $\alpha > \frac{d}{2}$ the existence of minimizers of (18) in $\mathcal{H}^\alpha(\Omega)$ is established by the direct method of the calculus of variations using the convexity of J_p and the fact that, by Lemma 2.4, $\mathcal{H}^\alpha$ continuously embeds into a set of Hölder continuous functions.

For $\alpha > \frac{d}{2}$ we believe that the minimizers of J_p^n of **Labelling Model 2** converge to minimizers of (18) in an appropriate regime, but the situation is more complicated than for **Labelling Model 1**: under **Labelling Model 2** (5) is no longer a sufficient condition on the scaling of ε with n for the convergence to hold. Thus if $\varepsilon \to 0$ too slowly the problem degenerates. In particular in the following theorem we identify the asymptotic behavior of minimizers of J_p both when $\alpha < \frac{d}{2}$, and if $\alpha > \frac{d}{2}$ but $\varepsilon \to 0$ too slowly.

The proof of the following may be found in section 7.6. The theorem is similar in spirit to Proposition 2.2(ii) in [39] where a similar phenomenon was discussed for the p-Laplacian regularized semi-supervised learning. We also mention that the PDE approach to a closely related p-Laplacian problem was recently introduced by Calder [40].

Theorem 4.3. *Let the assumptions of* **Labelling Model 2**, *and Theorem 2.2 hold. If $\alpha > \frac{d}{2}$, $\tau > 0$, and*

$$\varepsilon_n n^{\frac{1}{2\alpha}} \to \infty \qquad \text{as } n \to \infty \tag{20}$$

or if $\alpha < \frac{d}{2}$ then, with probability one, the sequence of minimizers v_n of $\mathsf{J}_\mathrm{p}^{(n)}$ converge to 0 in TL^2 as $n \to \infty$. That is, the minimizers of $\mathsf{J}_\mathrm{p}^{(n)}$ converge to the minimizer of $J_\infty^{(\alpha,\tau)}$ with the information about the labels being lost in the limit.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Remark 4.4. *We believe, but do not have a proof, that for* $\alpha > \frac{d}{2}$ *and* $\tau > 0$*, if*

$$\varepsilon_n n^{\frac{1}{2\alpha}} \to 0 \qquad \textit{as } n \to \infty$$

then, with probability one, any sequence of minimizers v_n *of* $\mathsf{J}_{\mathrm{p}}^{(n)}$ *is sequentially compact in* TL^2 *with* $\lim_{n\to\infty} \mathsf{J}_{\mathrm{p}}^{(n)}(v_n) = \min_{v \in L^2_\mu} \mathsf{J}_{\mathrm{p}}(v)$ *given by* (18), (19)*. If this holds then, under* **Labelling Model 2**, $\mathsf{J}_{\mathrm{p}}^{(n)}(u)$ *converges in an appropriate sense to a limiting objective function* $\mathsf{J}_{\mathrm{p}}(u)$*. Our numerical results support this conjecture.*

It is also of interest to consider the limiting probability distributions which arise under the two labelling models. Under **Labelling Model 2** *this density has, in physicist's notation, "Lebesgue density"* $\exp\big(-\mathsf{J}_{\mathrm{p}}(u)\big)$*. Under* **Labelling Model 1**, *however, we have shown that* $\mathsf{J}_{\mathrm{p}}^{(n)}(u)$ *converges in an appropriate sense to a limiting objective function* $\mathsf{J}_{\mathrm{p}}(u)$ *implying that (again in physicist's notation)* $\exp\big(-r_n^{-1}\mathsf{J}_{\mathrm{p}}^{(n)}(u)\big) \approx \exp\big(-n\mathsf{J}_{\mathrm{p}}(u)\big)$*. Thus under* **Labelling Model 1** *the posterior probability concentrates on a Dirac measure at the minimizer of* $\mathsf{J}_{\mathrm{p}}(u)$*.*

Based on this remark, the natural continuum probability limit concerns **Labelling Model 2**. The posterior probability is then given by

$$\nu_{\mathrm{p},2}(\mathrm{d}u) = \frac{1}{Z_{\mathrm{p},2}} e^{-\Phi_{\mathrm{p},2}(u;\gamma)} \nu_0(\mathrm{d}u) \tag{21}$$

where ν_0 is the centred Gaussian with covariance $\mathcal{C}$ given in Theorem 2.5 and $\Phi_{\mathrm{p},2}$ is given by (19). Since we require pointwise evaluation to make sense of $\Phi_{\mathrm{p},2}(u;\gamma)$ we, in general, require $\alpha > d$; however Proposition 2.6 gives conditions under which $\alpha > \frac{d}{2}$ will suffice. We will also consider the probability measure $\nu_{\mathrm{p},1}$ defined by

$$\nu_{\mathrm{p},1}(\mathrm{d}u) = \frac{1}{Z_{\mathrm{p},1}} e^{-\Phi_{\mathrm{p},1}(u;\gamma)} \nu_0(\mathrm{d}u) \tag{22}$$

where $\Phi_{\mathrm{p},1}$ is given by (16). The function $\Phi_{\mathrm{p},1}(u;\gamma)$ is defined in an L^2_μ sense and thus we require only $\alpha > \frac{d}{2}$ – see Theorem 2.5. Note, however, that this is not the limiting probability distribution that we expect for **Labelling Model 1** with the parameter choices leading to Theorem 4.2 since the argument above

ACCEPTED MANUSCRIPT

suggests that this will concentrate on a Dirac. However we include the measure $\nu_{\mathrm{p},1}$ in our discussions because, as we will show, it coincides with the analogous Bayesian level set measure $\nu_{\mathrm{ls},1}$ (defined below) in the small observational noise limit. Since $\nu_{\mathrm{ls},1}$ can be obtained by a natural scaling of the graph algorithm, which does not concentrate on Dirac, the relationship between $\nu_{\mathrm{p},1}$ and $\nu_{\mathrm{ls},1}$ is of interest as they are both, for small noise, relaxations of the same limiting object.

4.2. Bayesian Level Set

We now study probabilistic analogues of the Bayesian level set method, again using the measure ν_0 which is the centred Gaussian with covariance $\mathcal{C}$ given in Theorem 2.5 for some $\alpha > \frac{d}{2}$. Note that, from equation (13), for **Labelling Model 1**,

$$\begin{aligned} r_n \Phi_{\mathrm{ls}}^{(n)}(u;\gamma) &= \frac{1}{2\gamma^2}\frac{1}{n}\sum_{j\in Z'} \left|y(x_j) - S\big(u(x_j)\big)\right|^2 \\ &\approx \int_{\Omega'} \frac{1}{2\gamma^2}\left|y(x) - S\big(u(x)\big)\right|^2 \mathrm{d}\mu(x) \\ &:= \Phi_{\mathrm{ls},1}(u;\gamma) \end{aligned}$$

by a law of large numbers type argument of the type underlying the proof of Theorem 4.2.

Recall that, from the discussion following Proposition 3.2, this scaling corresponds to employing the finite dimensional Bayesian level set model with observational variance $\gamma^2 n$ so that the variance per observation is constant. Then the natural limiting probability measure is, in physicists notation, $\exp\big(-\mathsf{J}_{\mathrm{ls}}(u)\big)$ where

$$\mathsf{J}_{\mathrm{ls}}(u) = J_\infty^{(\alpha,\tau)}(u) + \Phi_{\mathrm{ls},1}(u;\gamma).$$

Expressed in terms of densities with respect to the Gaussian prior this gives

$$\nu_{\mathrm{ls},1}(\mathrm{d}u) = \frac{1}{Z_{\mathrm{ls},1}} e^{-\Phi_{\mathrm{ls},1}(u;\gamma)}\nu_0(\mathrm{d}u). \tag{23}$$

Since $\Phi_{\mathrm{ls},1}(u;\gamma)$ makes sense in L^2_μ we require only $\alpha > \frac{d}{2}$. The measure $\nu_{\mathrm{ls},1}$ is the natural analogue of the finite dimensional measure $\nu_{\mathrm{ls}}^{(n)}$ under this label

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

model. Under **Labelling Model 2** we take $r_n = 1$. We obtain a measure $\nu_{\text{ls},2}$ in the form (23) found by replacing $\nu_{\text{ls},1}$ by $\nu_{\text{ls},2}$ and $\Phi_{\text{ls},1}$ by

$$\Phi_{\text{ls},2}(u;\gamma) := \sum_{j\in Z'} \frac{1}{2\gamma^2}\big|y(x_j) - S\big(u(x_j)\big)\big|^2. \tag{24}$$

In this case the observational variance is not-rescaled by n since the total number of labels is fixed. Since we require pointwise evaluation to make sense of $\Phi_{\text{ls},2}(u;\gamma)$ we, in general, require $\alpha > d$; however Proposition 2.6 gives conditions under which $\alpha > \frac{d}{2}$ will suffice.

Remark 4.5. *Note that* $\mathsf{J}_{\text{ls}}^{(n)}$ *and* J_{ls} *cannot be connected via* Γ*-convergence. Indeed, if* $\mathsf{J}_{\text{ls}} = \Gamma\text{-}\lim_{n\to\infty} \mathsf{J}_{\text{ls}}^{(n)}$ *then* J_{ls} *would be lower semi-continuous [41]. When* $\tau > 0$ *compactness of minimizers follows directly from the compactness property of the quadratic forms* $J_n^{(\alpha,\tau)}$*, see Theorem 2.2. Now since compactness of minimizers plus lower semi-continuity implies existence of minimizers then the above reasoning implies there exists minimizers of* J_{ls}*. But as in the discrete case, Proposition 3.2, multiplying any* u *by a constant less than one leads to a smaller value of* J_{ls}*. Hence the infimum cannot be achieved. It follows that* $\mathsf{J}_{\text{ls}} \neq \Gamma\text{-}\lim_{n\to\infty} \mathsf{J}_{\text{ls}}^{(n)}$.

4.3. Small Noise Limit

As for the finite graph problems, the labeled data can be viewed as arising from different generative models. In the probit formulation, the generative models for the labels are given by

$$y(x) = S(u(x) + \xi(x)), \quad \xi \sim N(0,\gamma^2 I),$$
$$y(x_j) = S(u(x_j) + \xi_j), \quad \xi_j \overset{\text{iid}}{\sim} N(0,\gamma^2),$$

for **Labelling Model 1**, **Labelling Model 2** respectively; S is the sign function. The functionals $\Phi_{\text{p},1}$, $\Phi_{\text{p},2}$ then arise as the negative log-likelihoods from these models. Similarly, in the Bayesian level set formulation the generative models are given by

$$y(x) = S(u(x)) + \xi(x), \quad \xi \sim N(0,\gamma^2 I),$$
$$y(x_j) = S(u(x_j)) + \xi_j, \quad \xi_j \overset{\text{iid}}{\sim} N(0,\gamma^2).$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

leading to the functionals $\Phi_{\mathrm{ls},1}$, $\Phi_{\mathrm{ls},2}$.

We show that in the zero noise limit the Bayesian level set and probit posterior distributions coincide. However for $\gamma > 0$ they differ: note, for example, that the probit model enforces binary data, whereas the Bayesian level set model does not. It has been observed that the Bayesian level set posterior can be used to produce similar quality classification to the Ginzburg-Landau posterior, at significantly lower computational cost [42]. The small noise limit is important for two reasons: firstly in many applications labelling is very accurate and considering the zero noise limit is therefore instructive; secondly recent work [43] shows that the zero noise limit provides useful information about the efficiency of algorithms applied to sample the posterior distribution and, in particular, constants derived from the zero noise limit appear in lower bounds on average acceptance probability and mean square jump in such algorithms.

Proof of the following is given in section 7.7.

Theorem 4.6.

(i) *Let Assumptions 2–3 hold, and assume that $\alpha > d$. Let the assumptions of* **Labelling Model 1** *hold. Define the set*

$$B_{\infty,1} = \{u \in C(\Omega;\mathbb{R}) \,|\, y(x)u(x) > 0 \text{ for a.e. } x \in \Omega'\}$$

and the probability measure

$$\nu_1(\mathrm{d}u) = \mathsf{Z}^{-1}\mathbb{1}_{B_{\infty,1}}(u)\nu_0(\mathrm{d}u)$$

where $\mathsf{Z} = \nu_0(B_{\infty,1})$. Consider the posterior measures $\nu_{\mathrm{p},1}$ defined in (22) *and $\nu_{\mathrm{ls},1}$ defined in* (23)*. Then $\nu_{\mathrm{p},1} \Rightarrow \nu_1$ and $\nu_{\mathrm{ls},1} \Rightarrow \nu_1$ as $\gamma \to 0$.*

(ii) *Let Assumptions 2–3 hold, and assume that $\alpha > d$. Let the assumptions of* **Labelling Model 2** *hold. Define the set*

$$B_{\infty,2} = \{u \in C(\Omega;\mathbb{R}) \,|\, y(x_j)u(x_j) > 0 \text{ for each } j \in Z'\}$$

and the probability measure

$$\nu_2(\mathrm{d}u) = \mathsf{Z}^{-1}\mathbb{1}_{B_{\infty,2}}(u)\nu_0(\mathrm{d}u)$$

ACCEPTED MANUSCRIPT

where $\mathsf{Z} = \nu_0(B_{\infty,2})$. *Then* $\nu_{\mathrm{p},2} \Rightarrow \nu_2$ *and* $\nu_{\mathrm{ls},2} \Rightarrow \nu_2$ *as* $\gamma \to 0$.

Remark 4.7. *The assumption that* $\alpha > d$ *in both parts of the above theorem can be relaxed to* $\alpha > d/2$ *if the conclusions of Proposition 2.6 are satisfied.*

4.4. Kriging

One can define kriging in the continuum setting [12] analogously to the discrete setting; we consider this numerically in section 5. In the case of **Labelling Model 2**, the limiting problem is to

$$\text{minimize } \mathsf{J}_{\mathrm{k}}(u) \coloneqq J_{\infty}^{(\alpha,\tau)}(u) \text{ subject to } u(x_j) = y_j \text{ for all } j \in Z'.$$

Kriging may also be defined for **Labelling Model 1** and without the hard constraint in the continuum setting, but we do not discuss either of these scenarios here.

5. Numerical Illustrations

In this section we describe the results of numerical experiments which illustrate or extend the developments in the preceding sections. In section 5.1 we study the effect of the geometry of the data on the classification problem, by studying an illustrative example in dimension $d = 2$. Section 5.2 studies how the relationship between the length-scale ϵ and the graph size n affects limiting behaviour. In section 5.3 we study graph based kriging. Finally, in section 5.4, we study continuum problems from the Bayesian perspective, studying the quantification of uncertainty in the resulting classification.

5.1. Effect of Data Geometry on Classification

We study how the geometry of the data affects the classification under **Labelling Model 1**, using the continuum probit model. Let $\Omega = (0,1)^2$. We first consider a uniform distribution ρ on the domain, and choose Ω_+, Ω_- to be balls of radius 0.05 centred at (0.25,0.25), (0.75,0.75) respectively. The decision boundary is then naturally the perpendicular bisector of the line segment joining the centers of these balls. We then modify ρ by introducing a channel of

ACCEPTED MANUSCRIPT

increasing depth in ρ dividing the domain in two vertically, and look at how this affects the decision boundary. Specifically, given $h \in [0, 1]$ we define ρ_h to be constant in the y-direction, and assume the cross-sections in the x-direction are as shown in Figure 2, so that the channel has depth $1 - h$. In order to numerically estimate the continuum probit minimizers, we construct a finite-difference approximation to each $\mathcal{L}$ on a uniform grid of 65536 points, which then provides an approximation to $\mathcal{A}$. The objective function $\mathsf{J}_{\mathrm{p}}^{(\infty)}$ is then minimized numerically using the linearly-implicit gradient flow method described in [17], Algorithm 4.

We consider both the effect of the channel depth parameter h and the parameter α on the classification; we fix $\tau = 10$ and $\gamma = 0.01$. In Figure 3 we show the minimizers arising from 5 different choices of h and $\alpha = 1, 2, 3$. As the depth of the channel is increased, the minimizers begin to develop a jump along the channel. As α is increased, the minimizers become less localized around the labeled regions, and the jump along the channel becomes sharper as a result. Note that the scale of the minimizers decreases as α increases. This could formally be understood from a probabilistic point of view: under the prior we have $\mathbb{E}\|u\|_{L^2}^2 = \mathrm{Tr}(\mathcal{A}^{-1}) \asymp \tau^{-2\alpha}$, and so a similar scaling may be expected to hold for the MAP estimators. In Figure 4 we show the sign of each minimizer in Figure 3 to illustrate the resulting classifications. As the depth of the channel is increased, the decision boundary moves continuously from the diagonal to the vertical bisector of the domain, with the transitional boundaries appearing almost as a piecewise linear combination of both boundaries. We also see that, despite the minimizers themselves differing significantly for different α, the classifications are almost invariant with respect to α.

5.2. Localization Bounds for Kriging and Probit

We study how the rate affects convergence to the continuum limits when the localization parameter decreases and the number of data points n is increased. We consider **Labelling model 2** using both the kriging and probit models; this serves to illustrate the result of Theorem 4.3, motivate Remark 4.4, and provide

ACCEPTED MANUSCRIPT

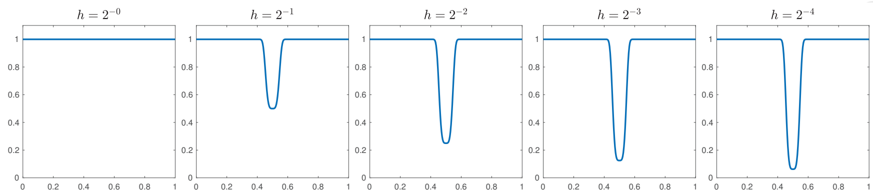

Figure 2: The cross sections of the data densities ρ_h we consider in subsection 5.1.

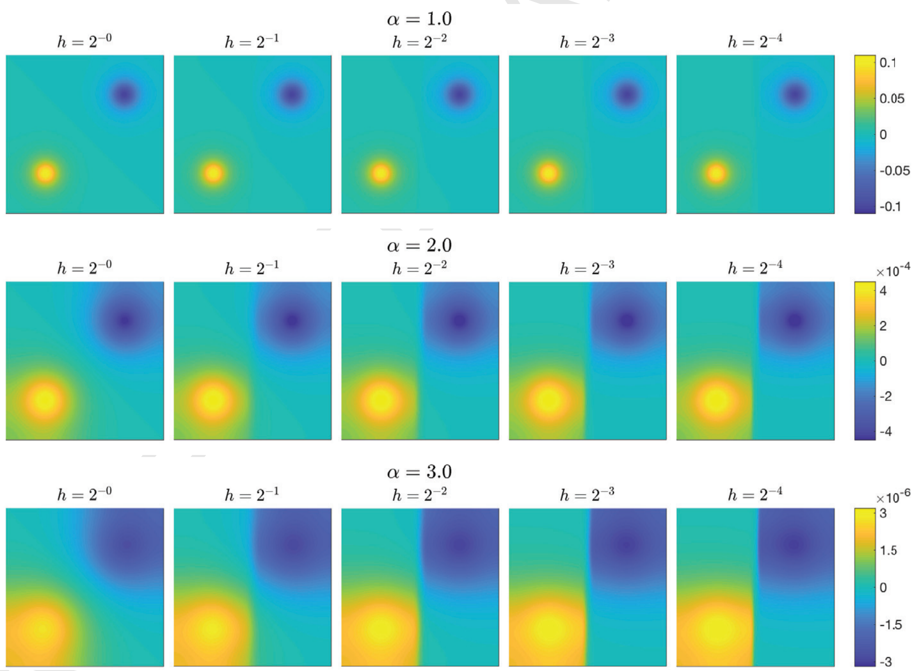

Figure 3: The minimizers of the functional $\mathsf{J}_{\mathrm{p}}^{(\infty)}$ for different values of h and α, as described in subsection 5.1.

ACCEPTED MANUSCRIPT

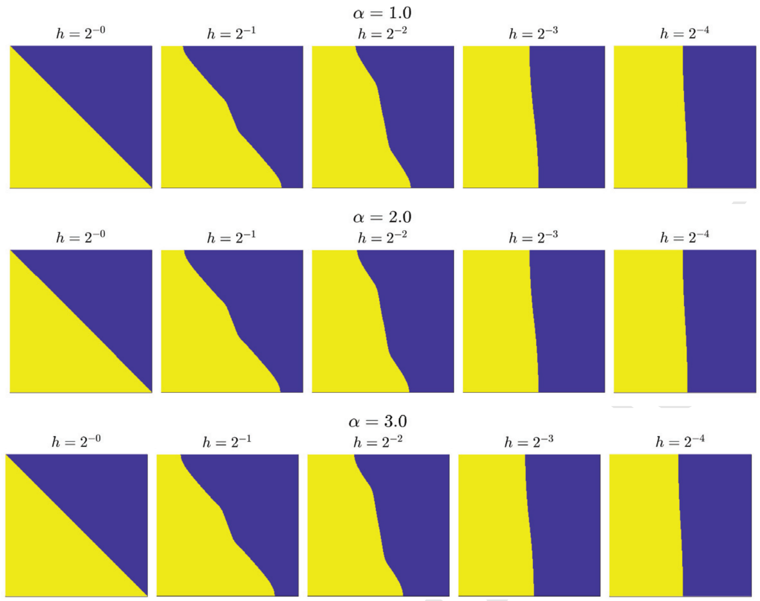

Figure 4: The sign of minimizers from Figure 3, showing the resulting classification.

a relation to the results of [39].

We work on the domain $\Omega = (0,1)^2$ and take a uniform data distribution ρ. In all cases we fix two datapoints which we label with opposite signs, and sample the remaining $n-2$ datapoints. For kriging we consider the situation where the data is viewed as noise-free so that the label values are interpolated. We calculate the minimizer u_n of $\mathsf{J}_{\mathrm{k}}^{(n)}$ numerically via the closed form solution

$$u_n = A^{(n),-1} R^* (R A^{(n),-1} R^*)^{-1} y,$$

where $R \in \mathbb{R}^{2\times n}$ is the mapping taking vectors to their values at the labeled points. In order to numerically estimate the continuum minimizer u of $\mathsf{J}_{\mathrm{k}}^{(\infty)}$, we construct a finite-difference approximation to $\mathcal{L}$ on a uniform grid of 65536 points. This leads to an approximation $\hat{\mathcal{A}}$ to $\mathcal{A}$, from which we again use the closed form solution to compute $\hat{u} \approx u$:

$$\hat{u} = \hat{\mathcal{A}}^{-1} \hat{R}^* (\hat{R} \hat{\mathcal{A}}^{-1} \hat{R}^*)^{-1} y,$$

where $\hat{R} \in \mathbb{R}^{2\times 65556}$ takes discrete functions to their values at the labeled points.

ACCEPTED MANUSCRIPT

In Figure 5 (left) we show how the $L^2_{\mu_n}$ error between u_n and $\hat{u}$ varies with respect to ε for increasing values of n. All errors are averaged over 200 realizations of the unlabeled datapoints, and we consider 100 uniformly spaced values of ε between 0.005 and 0.5. We see that ε must belong to a 'sweet-spot' in order to make the error small – if ε is too small or too large convergence doesn't occur. The right hand side of the figure shows how these lower and upper bounds vary with n; the bounds are defined numerically as the points where the second derivative of the error curve changes sign. The rates are in agreement with the results and conjectures up to logarithmic terms, although the sharp bounds are not obtained – we see that the lower bounds are larger than $\mathcal{O}(n^{-\frac{1}{2}})$, and the upper bounds are smaller than $\mathcal{O}(n^{-\frac{1}{2\alpha}})$. It is possible that the sharp bounds may be approached in a more asymptotic (and computationally infeasible) regime.

Similarly, we note that the minimum error for $\alpha = 2$ in Figure 5 decreases very slowly in the range of n we considered. This again indicates that we are not yet in the asymptotic regime at $n = 1600$. Further experiments (not included) for larger values of n show that the minimum error does converge as $n \to \infty$ as expected.

For the probit model we take $\gamma = 0.01$ and use the same gradient flow algorithm as in subsection 5.1 for both the continuum and discrete minimizers. Figure 6 shows the errors, analogously to Figure 5. Note that the errors are plotted on logarithmic axes here, as unlike the kriging minimizers, there is no restriction for the minimizers to be on the same scale as the labels. We see that the same trend is observed in terms of requiring upper and lower bounds on ε, and a shift of the error curves towards the left as n is increased.

5.3. Extrapolation on Graphs

We consider the problem of smoothly extending a sparsely defined function on a graph to the entire graph. Such extrapolation was studied in [44], and was achieved via the use of a weighted nonlocal Laplacian. We use the kriging model with **Labelling Model 2**, labelling two points with opposite signs, and setting $\gamma = 0$. We fix a set of datapoints $\{x_j\}_{j=1}^n$, $n = 1600$, drawn from the

ACCEPTED MANUSCRIPT

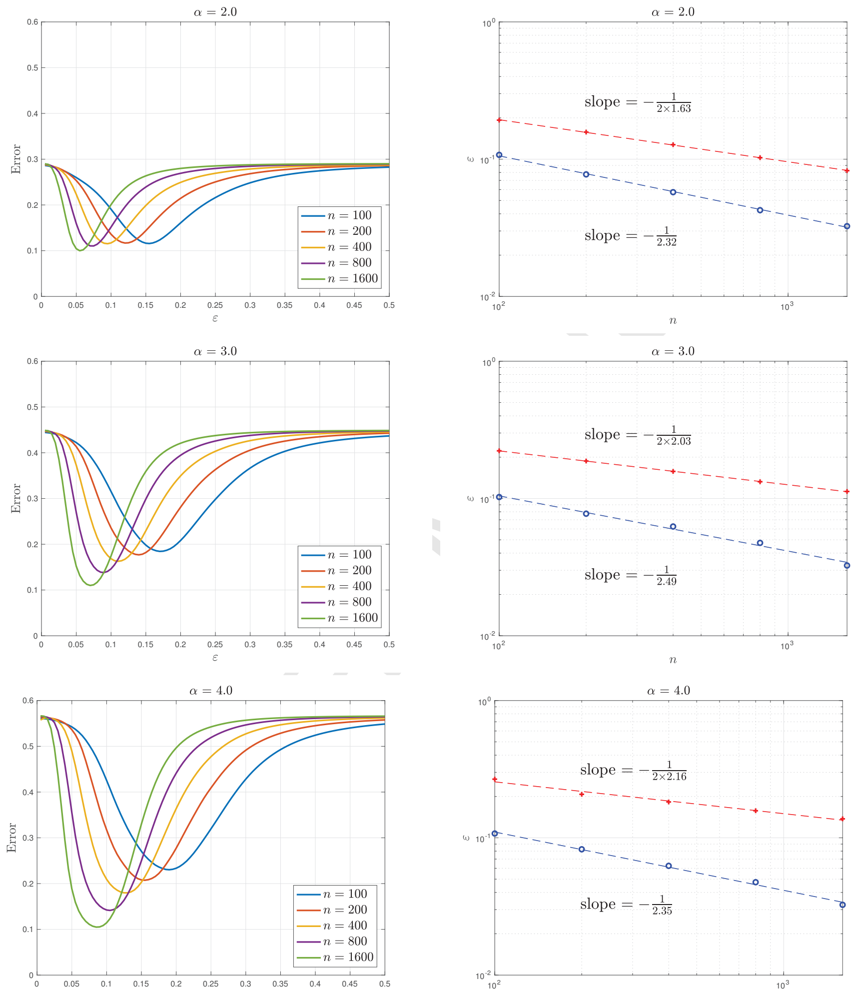

Figure 5: (Left) The $L^2_{\mu_n}$ error between discrete minimizers and continuum minimizers of the kriging model versus localization parameter ε, for different values of n. (Right) The upper and lower bounds for $\varepsilon(n)$ to provide convergence. The slopes of the lines of best fit provide estimates of the rates.

ACCEPTED MANUSCRIPT

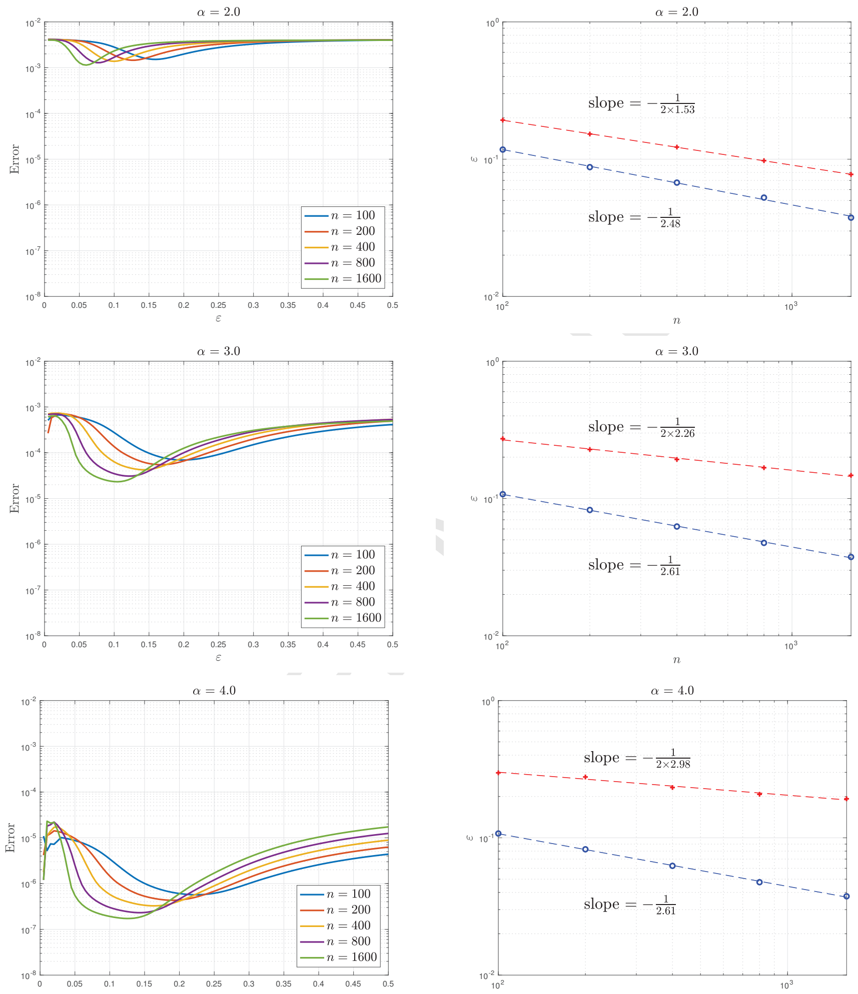

Figure 6: (Left) The $L^2_{\mu_n}$ error between discrete minimizers and continuum minimizers of the probit model versus localization parameter ε, for different values of n. (Right) The upper and lower bounds for $\varepsilon(n)$ to provide convergence. The slopes of the lines of best fit provide estimates of the rates.

ACCEPTED MANUSCRIPT

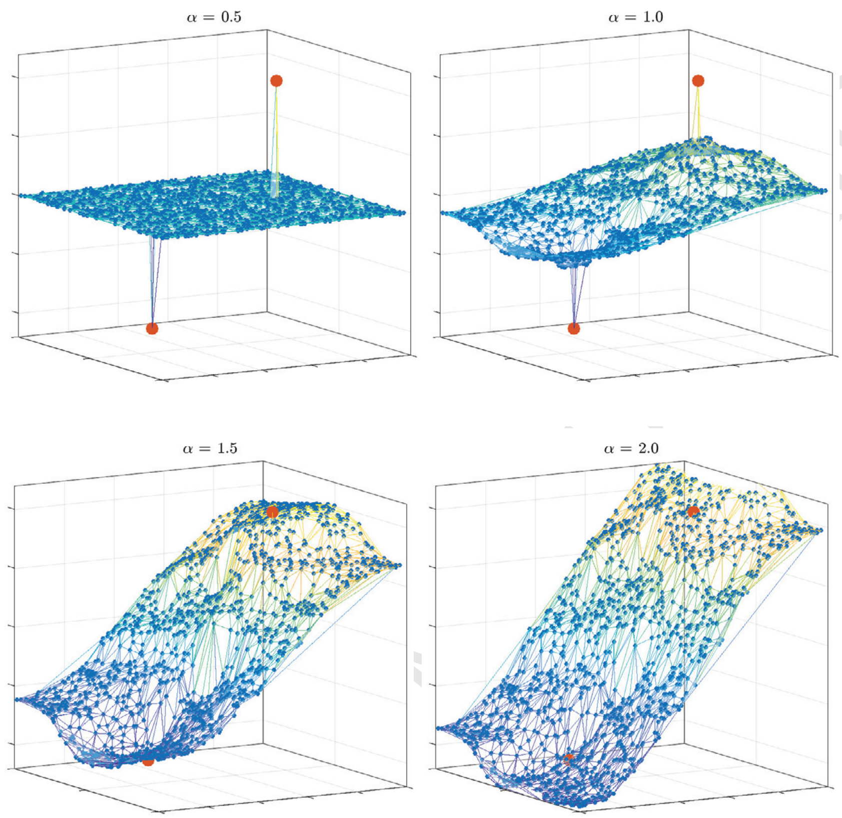

Figure 7: The extrapolation of a sparsely defined function on a graph using the kriging model, for various choices of parameter α.

uniform density on the domain $\Omega = (0,1)^2$. We fix $\tau = 1$ and look at how the smoothness of minimizers of the kriging functional $\mathsf{J}_{\mathrm{k}}^{(n)}$ varies with α. The minimizers are computed directly from the closed form solution, as in subsection 5.2. When $\alpha > d/2$ we choose ε to approximately minimize the $L^2_{\mu_n}$ errors between the discrete and continuum solutions (since the continuum solution is non-trivial). When $\alpha \leq d/2$ a representative ε is chosen which is approximately twice the connectivity radius. The minimizers are shown in Figure 7 for $\alpha = 0.5, 1.0, 1.5, 2.0$. Spikes are clearly visible for $\alpha \leq d/2 = 1$: the requirement for $\alpha > d/2$ to avoid spikes appears to be essential.

ACCEPTED MANUSCRIPT

5.4. Bayesian Level Set for Sampling

We now turn to the problem of sampling the conditioned continuum measures introduced in subsections 4.1 and 4.2, specifically their common $\gamma \to 0$ limit. From this sampling we can, for example, calculate the mean of the classification, which may be used to define a measure of uncertainty of the classification at each point. This is because, for binary random variables, the mean determines the variance. Knowing the uncertainty in classification has great potential utility, for example in active learning in guiding where to place resources in labelling in order to reduce uncertainty.

We fix $\Omega = (0,1)^2$. The data distribution ρ is shown in Figure 8; it is constructed as a continuum analogue of the two moons distribution [45], with the majority of its mass concentrated on two curves. The contrast ratio in the sampling density ρ is approximately 100:1 between the values on and off of the curves. The resulting operator $\mathcal{L}$ contains significant clustering information: in Figure 8 we show the second eigenfunction of $\mathcal{L}$, termed the Fiedler vector in analogy with second eigenvector of the graph Laplacian. The sign of this function provides a good estimate for the decision boundary in an unsupervised context. We use **Labelling Model 2**, labelling a single point on each curve with opposing signs as indicated by $\bullet$ and $\circ$ in Figure 8.

Sampling is performed using the preconditioned Crank-Nicolson MCMC algorithm [46], which has favourable dimension-independent statistical properties, as demonstrated in [30] in the graph-based setting of relevance here. We consider three choices of $\alpha > d/2$, and two choices of inverse length-scale parameter τ. In general we require $\alpha > d$ for the measure ν_2 in Theorem 4.6 to be well-defined. However numerical evidence suggests that the conclusions of Proposition 2.6 are satisfied with this choice of ρ, implying that we may make use of Remark 4.7 and that $\alpha > \frac{d}{2}$ suffices. The operator $\mathcal{L}$ is discretized using a finite difference method on a square grid of 40000 points, and sampling is performed on the span of its first 500 eigenfunctions.

In Figure 9 we show the mean of the sign of samples on the left hand side, for each choice of α, after fixing $\tau = 1$. Note that uncertainty is greater the further

ACCEPTED MANUSCRIPT

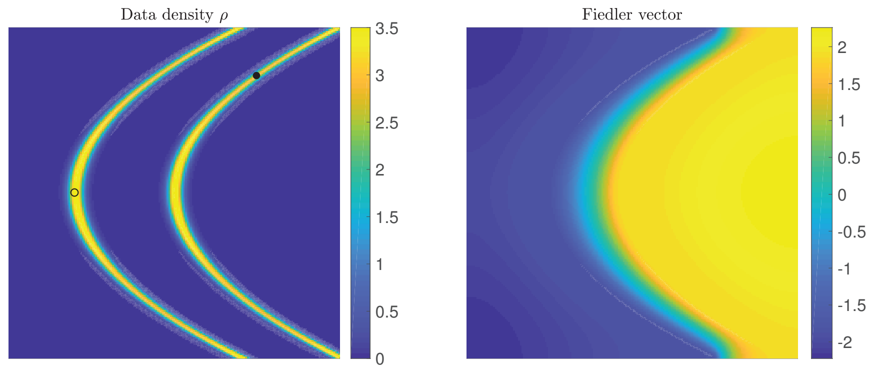

Figure 8: (Left) The data distribution ρ used in the MCMC experiments, and the locations of the two labeled datapoints. (Right) The second eigenfunction of the operator $\mathcal{L}$ corresponding to ρ.

the values of the mean are from ± 1: specifically we have that $\mathrm{Var}\big(S(u(x)\big) = 1 - \big[\mathbb{E}(S(u(x)))\big]^2$. We see that the classification on the curves where the data concentrates is fairly certain, whereas classification away from the curves is uncertain; furthermore the certainty increases away from the curves slightly as α is increased. Samples $S(u)$ are also shown in the same figure; the uncertainty away from the curves is illustrated also by these samples.

In Figure 10 we show the same results, but with the choice $\tau = 0.2$ so that samples possess a longer length scale. The classification certainty now propagates away from the curves more easily. The effect of the asymmetry of the labelling is also visible in the mean for the case $\alpha = 4$: uncertainty is higher in the bottom-left corner than the top-left corner.

Since the prior on the latent random field u may be difficult to ascertain in applications, the sensitivity of the classification on the choice of the parameters α, τ indicates that it could be wise to employ hierarchical Bayesian methods to learn appropriate values for them along with the latent field u. Dimension robust MCMC methods are available to sample such hierarchical distributions [47], and application to classification problems are shown in that paper.

ACCEPTED MANUSCRIPT

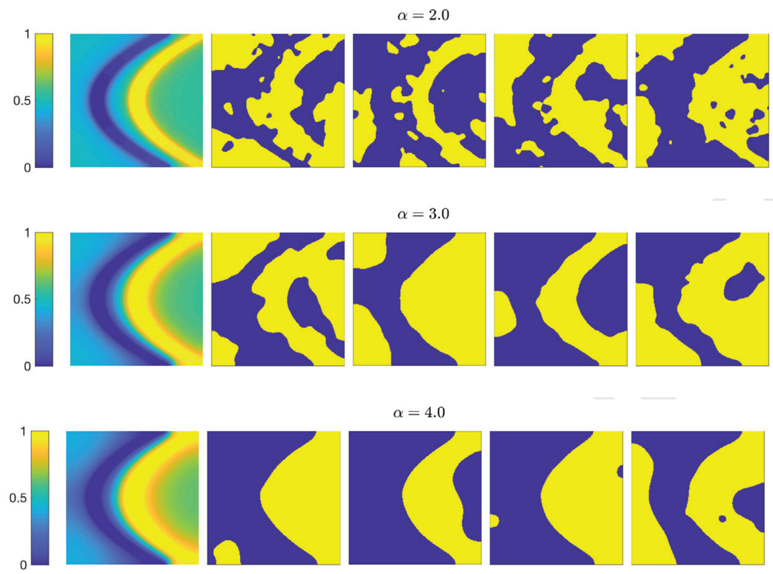

Figure 9: (Left) The mean $\mathbb{E}\big(S(u)\big)$ of the classification arising from the conditioned measure ν_2. (Right) Examples of samples $S(u)$ where $u \sim \nu_2$. Here we choose $\tau = 1$.

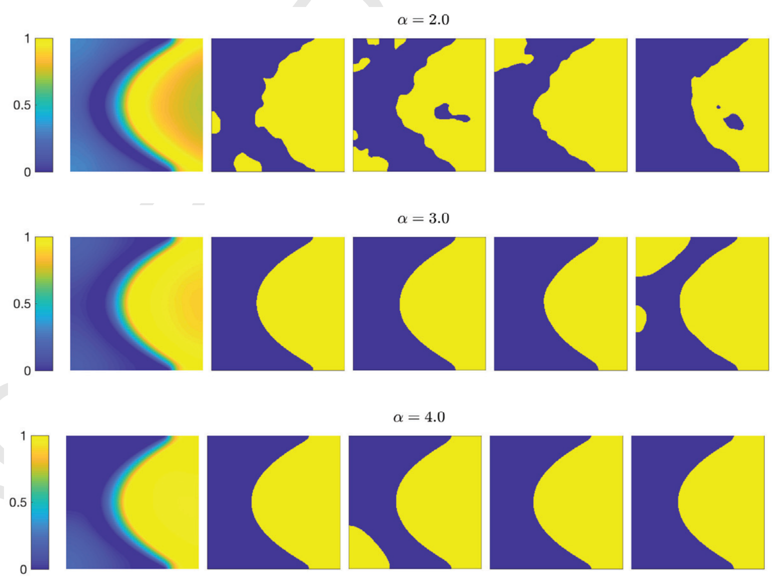

Figure 10: (Left) The mean $\mathbb{E}\big(S(u)\big)$ of the classification arising from the conditioned measure ν_2. (Right) Examples of samples $S(u)$ where $u \sim \nu_2$. Here we choose $\tau = 0.2$.

ACCEPTED MANUSCRIPT

6. Conclusions

In this paper we have studied large graph limits of semi-supervised learning problems in which smoothness is imposed via a shifted graph Laplacian, raised to a power. Both optimization and Bayesian approaches have been considered. To keep the exposition manageable in length we have confined our attention to the unnormalized graph Laplacian. However, one may instead choose to work with the normalized graph Laplacian $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, in place of $L = D - W$. In the normalized case the continuum PDE operator is given by

$$\mathcal{L}u = -\frac{1}{\rho^{3/2}} \nabla \cdot \left(\rho^2 \nabla \Big(\frac{u}{\rho^{1/2}} \Big) \right)$$

with no flux boundary conditions: $\nabla \Big(\frac{u}{\rho^{1/2}} \Big) \cdot \nu = 0$ on $\partial\Omega$, where ν is the outside unit normal vector to $\partial\Omega$. Theorems 2.2, 4.2 and 4.6 generalize in a straightforward way to such a change in the graph Laplacian.

Future directions stemming from the work in this paper include: (i) providing a limit theorem for probit MAP estimators under **Labelling Model 2**; (ii) providing limit theorems for the Bayesian probability distributions considered, using the machinery introduced in [30, 29]; (iii) using the limiting problems in order to analyze and quantify efficiency of algorithms on large graphs; (iv) invoking specific sources of data and studying the effectiveness of PDE limits in comparison to non-local limits.

Acknowledgements The authors are grateful to Ian Tice and Giovanni Leoni for valuable insights and references. The authors are thankful to Christopher Sogge and Steve Zelditch for useful background informtion. The authors are also grateful to the Center for Nonlinear Analysis (CNA) and Ki-Net (NSF Grant RNMS11-07444). MT is grateful to the Cantab Capital Institute for the Mathematics of Information (CCIMI) and the Cambridge Image Analysis (CIA) group. DS acknowledges the support of the National Science Foundation under the grant DMS 1516677 and DMS 1814991. MMD and AMS are supported by AFOSR Grant FA9550-17-1-0185 and the National Science Foundation grant

ACCEPTED MANUSCRIPT

DMS 1818977. DS and MT acknowledge funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska–Curie grant agreement No 777826.

[1] X. Zhu, Semi-supervised learning literature survey, Tech. rep., Computer Science, University of Wisconsin-Madison (2005).

[2] X. Zhu, Semi-supervised learning with graphs, Ph.D. thesis, Carnegie Mellon University, language technologies institute, school of computer science (2005).

[3] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in neural information processing systems, 2002, pp. 585–591 (2002).

[4] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in neural information processing systems, 2002, pp. 849–856 (2002).

[5] U. Von Luxburg, A tutorial on spectral clustering, Statistics and computing 17 (4) (2007) 395–416 (2007).

[6] A. Blum, S. Chawla, Learning from labeled and unlabeled data using graph mincuts, Tech. rep., CMU Tech Report (2001).

[7] Y. Boykov, O. Veksler, R. Zabih, Fast approximate energy minimization via graph cuts, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (11) (2001) 1222–1239 (2001).

[8] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Transactions on pattern analysis and machine intelligence 22 (8) (2000) 888–905 (2000).

[9] A. Madry, Fast approximation algorithms for cut-based problems in undirected graphs, in: Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on, IEEE, 2010, pp. 245–254 (2010).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[10] S. Z. Li, Markov random field modeling in computer vision, Springer Science & Business Media, 2012 (2012).

[11] A. Szlam, X. Bresson, Total variation and Cheeger cuts, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 1039–1046 (2010).

[12] G. Wahba, Spline models for observational data, SIAM, 1990 (1990).

[13] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using Gaussian fields and harmonic functions, in: Proceedings of the 20th International Conference on Machine Learning, Vol. 3, 2003, pp. 912–919 (2003).

[14] R. Neal, Regression and classification using Gaussian process priors, Bayesian Statistics 6 (1998) 475 (1998).

[15] C. K. Williams, C. E. Rasmussen, Gaussian processes for regression, in: Advances in neural information processing systems, 1996, pp. 514–520 (1996).

[16] C. Bishop, Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn, Springer, New York (2007).

[17] A. L. Bertozzi, X. Luo, A. M. Stuart, K. C. Zygalakis, Uncertainty quantification in the classification of high dimensional data, arXiv preprint arXiv:1703.08816 (2017).

[18] A. L. Bertozzi, A. Flenner, Diffuse interface models on graphs for classification of high dimensional data, Multiscale Modeling & Simulation 10 (3) (2012) 1090–1118 (2012).

[19] R. Cristoferi, M. Thorpe, Large data limit for a phase transition model with the p-Laplacian on point clouds, arxiv preprint arXiv:1802.08703 (2018).

[20] M. Thorpe, F. Theil, Asymptotic analysis of the Ginzburg-Landau functional on point clouds, to appear in the Proceedings of the Royal Society

ACCEPTED MANUSCRIPT

of Edinburgh Section A: Mathematics, arXiv preprint arXiv:1604.04930 (2017).

[21] Y. Van Gennip, A. L. Bertozzi, Γ-convergence of graph Ginzburg-Landau functionals, Advances in Differential Equations 17 (11-12) (2012) 1115–1180 (2012).

[22] M. Burger, S. Osher, A survey on level set methods for inverse problems and optimal design, Europ. J. Appl. Math. 16 (2005) 263–301 (2005).

[23] M. A. Iglesias, Y. Lu, A. M. Stuart, A Bayesian level set method for geometric inverse problems, Interfaces and Free Boundary Problems (2015).

[24] U. Von Luxburg, M. Belkin, O. Bousquet, Consistency of spectral clustering, The Annals of Statistics (2008) 555–586 (2008).

[25] N. García Trillos, D. Slepčev, A variational approach to the consistency of spectral clustering, Applied and Computational Harmonic Analysis (2016).

[26] B. Nadler, N. Srebro, X. Zhou, Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data, in: Advances in neural information processing systems, 2009, pp. 1330–1338 (2009).

[27] X. Zhou, M. Belkin, Semi-supervised learning by higher order regularization., in: AISTATS, 2011, pp. 892–900 (2011).

[28] X. Zhu, J. D. Lafferty, Z. Ghahramani, Semi-supervised learning: From Gaussian fields to Gaussian processes, Tech. rep., CMU Tech Report:CMU-CS-03-175 (2003).

[29] N. García Trillos, D. Sanz-Alonso, Continuum limit of posteriors in graph Bayesian inverse problems, arXiv preprint arXiv:1706.07193 (2017).

[30] N. García Trillos, Z. Kaplan, T. Samakhoana, D. Sanz-Alonso, On the consistency of graph-based Bayesian learning and the scalability of sampling algorithms, arXiv preprint arXiv:1710.07702 (2017).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

[31] L. Hörmander, The spectral function of an elliptic operator, Acta Math 121 (1968) 193–218 (1968).

[32] H. Abels, Short lecture notes: Interpolation theory and function spaces, `http://www.uni-r.de/Fakultaeten/nat_Fak_I/abels/SkriptInterpolationstheorieSoSe11.pdf` (2011).

[33] J. Peetre, On an interpolation theorem of Foiaş and Lions, Acta Sci. Math. (Szeged) 25 (1964) 255–261 (1964).

[34] J. E. Gilbert, Interpolation between weighted L^p-spaces, Ark. Mat. 10 (1972) 235–249 (1972).

[35] M. Dashti, A. M. Stuart, The Bayesian approach to inverse problems, in: Handbook of Uncertainty Quantification, Springer, 2016, p. arxiv preprint arXiv:1302.6989 (2016).

[36] D. Grieser, Uniform bounds for eigenfunctions of the Laplacian on manifolds with boundary, Comm. Partial Differential Equations 27 (7-8) (2002) 1283–1299 (2002).

[37] C. D. Sogge, S. Zelditch, Riemannian manifolds with maximal eigenfunction growth, Duke Math. J. 114 (3) (2002) 387–437 (2002).

[38] G. Leoni, A first course in Sobolev spaces, 2nd Edition, Vol. 181 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2017 (2017).

[39] D. Slepčev, M. Thorpe, Analysis of p-Laplacian regularization in semi-supervised learning, arXiv preprint arXiv:1707.06213 (2017).

[40] J. Calder, The game theoretic p-Laplacian and semi-supervised learning with few labels, arXiv preprint arXiv:1711.10144 (2017).

[41] A. Braides, Γ-Convergence for Beginners, Oxford University Press, Oxford, 2002 (2002).

ACCEPTED MANUSCRIPT

[42] M. Dunlop, C. Elliott, V. Hoang, A. Stuart, Reconciling Bayesian and total variation methods for binary inversion, arXiv preprint arXiv:1706.01960 (2017).

[43] A. L. Bertozzi, X. Luo, O. Papaspiliopoulos, A. M. Stuart, Scalable and robust sampling methods for Bayesian graph-based semi-supervised learning, In preparation (2018).

[44] Z. Shi, S. Osher, W. Zhu, Weighted nonlocal Laplacian on interpolation from sparse data, Journal of Scientific Computing 73 (2-3) (2017) 1164–1177 (2017).

[45] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Advances in neural information processing systems, 2004, pp. 321–328 (2004).

[46] S. L. Cotter, G. O. Roberts, A. M. Stuart, D. White, MCMC methods for functions: modifying old algorithms to make them faster., Statistical Science 28 (3) (2013) 424–446 (2013).

[47] V. Chen, M. M. Dunlop, O. Papasiliopoulos, A. M. Stuart, Robust MCMC sampling with non-Gaussian and hierarchical priors in high dimensions, arXiv preprint arXiv:1803.03344 (2018).

[48] P. Grisvard, Elliptic problems in nonsmooth domains, SIAM, 2011 (2011).

[49] N. García Trillos, D. Slepčev, Continuum limit of total variation on point clouds, Archive for Rational Mechanics and Analysis 220 (1) (2016) 193–241 (2016).

[50] M. Thorpe, A. M. Johansen, Convergence and rates for fixed-interval multiple-track smoothing using k-means type optimization, Electronic Journal of Statistics 10 (2) (2016) 3693–3722 (2016).

[51] M. Thorpe, F. Theil, A. M. Johansen, N. Cade, Convergence of the k-means minimization problem using Γ-convergence, SIAM Journal on Applied Mathematics 75 (6) (2015) 2444–2474 (2015).

ACCEPTED MANUSCRIPT

[52] G. Dal Maso, An Introduction to Γ-Convergence, Springer, 1993 (1993).

[53] N. García Trillos, D. Slepčev, On the rate of convergence of empirical measures in ∞-transportation distance, Canadian Journal of Mathematics 67 (2015) 1358–1383 (2015).

[54] M. Abramowitz, I. A. Stegun, Handbook of mathematical functions with formulas, graphs, and mathematical tables, Vol. 55 of National Bureau of Standards Applied Mathematics Series, For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964 (1964).

7. Appendix

7.1. Function Spaces

Here we establish the equivalence between the spectrally defined Sobolev spaces, $\mathcal{H}^s(\Omega)$ and the standard Sobolev spaces.

We denote by

$$H^2_N(\Omega) = \left\{ u \in H^2(\Omega) \,:\, \frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega \right\}$$

the domain of $\mathcal{L}$. Analogously we denote by $H^{2m}_N(\Omega)$ the domain of $\mathcal{L}^m$, that is

$$H^{2m}_N(\Omega) = \left\{ u \in H^{2m}(\Omega) \,:\, \frac{\partial \mathcal{L}^r u}{\partial n} = 0 \text{ for all } 0 \le r \le m-1 \text{ on } \partial\Omega \right\}$$

Finally we let $H^{2m+1}_N(\Omega) = H^{2m+1}(\Omega) \cap H^{2m}_N(\Omega)$.

For $m \ge 0$ and $u, v \in H^{2m+1}_N(\Omega)$ let $\langle u, v\rangle_{2m+1,\mu} = \int_\Omega \nabla \mathcal{L}^m u \cdot \nabla \mathcal{L}^m v \rho^2 dx$ and for $u, v \in H^{2m}_N(\Omega)$ let $\langle u, v\rangle_{2m,\mu} = \int_\Omega (\mathcal{L}^m u)(\mathcal{L}^m v)\rho \,\mathrm{d}x$. We note that on the L^2_μ orthogonal complement of the constant function 1, $\langle \cdot, \cdot \rangle_{2m+1,\mu}$ defines an inner product, which due to Poincaré inequality is equivalent to the standard inner product on $H^{2m+1}(\Omega)$. We also note that $\langle \varphi_k, \varphi_k\rangle_{2m+1,\mu} = \lambda_k^{2m+1}$, where we recall that φ_k is unit eigenvector of $\mathcal{L}$ corresponding to λ_k.

ACCEPTED MANUSCRIPT

Lemma 7.1. *Under Assumptions 2 - 3, for any integer* $s \geq 0$

$$H^s_N(\Omega) = \mathcal{H}^s(\Omega)$$

and the associated inner products $\langle \cdot, \cdot \rangle_{s,\mu}$ *and* $\langle\!\langle \cdot, \cdot \rangle\!\rangle_{s,\mu}$ *are equivalent on the* L^2_μ *orthogonal complement of the constant function.*

Proof. For $s = 0$, $H^0_N = L^2$ by definition and $\mathcal{H}^0 = L^2$ by the fact that $\{\varphi_k : k = 1, \dots\}$ is an orthonormal basis.

To show the claim for $s = 1$, we recall that $\int \nabla\varphi_k \cdot \nabla\varphi_j \rho^2 dx = \int \varphi_k \mathcal{L}\varphi_j \rho dx = \lambda_k \delta^j_k$. Therefore $\left\{\frac{\varphi_k}{\sqrt{\lambda_k}} : k \geq 1\right\}$ is an orthonormal basis of the orthogonal complement of the constant function, $1^\perp$, in H^1_N with respect to the inner product $(u, v) = \int \nabla_u \cdot \nabla v \rho^2 dx$ which is equivalent to the standard inner product of H^1_N on $1^\perp$. Since an expansion in the basis $\{\varphi_k\}_k$ is unique, this implies that for any $u \in H^1_N = H^1$ the series $\sum_k a_k \varphi_k$ converges in H^1 to u. Consequently if $u \in H^1_N$ then $\infty > \int |\nabla u|^2 \rho^2 dx = \int |\sum_k a_k \nabla\varphi_k|^2 \rho^2 dx = \sum_k a_k^2 \lambda_k$ which implies that $u \in \mathcal{H}^1$. So $H^1_N \subseteq \mathcal{H}^1$.

On the other hand, if $u \in \mathcal{H}^1$ then $u = \sum_k a_k \varphi_k$ with $\sum_k \lambda_k a_k^2 < \infty$. Therefore $u = \bar{u} + \sum_{k=2}^\infty a_k \sqrt{\lambda_k} \frac{\varphi_k}{\sqrt{\lambda_k}}$, where $\bar{u}$ is the average of u. Since $\frac{\varphi_k}{\sqrt{\lambda_k}}$ are orthonormal in scalar product with topology equivalent to H^1, the series converges in H^1. Therefore $u \in H^1 = H^1_N$.

Assume now that the claim holds for all integers less than s. We split the proof of the induction step into two cases:

Case 1° Consider s even; that is $s = 2m$ for some integer $m > 0$.

Assume $u \in H^{2m}_N$. Then $\nabla\mathcal{L}^r u \cdot \vec{n} = 0$ on $\partial\Omega$ for all $r < m$. By the induction hypothesis $\sum_k \lambda_k^{2m-1} a_k^2 < \infty$. Since $\mathcal{L}$ is a continuous operator from $\mathcal{H}^2$ to L^2 one obtains by induction that $\mathcal{L}^{m-1}u = \sum_k a_k \mathcal{L}^{m-1}\varphi_k = \sum a_k \lambda_k^{m-1}\varphi_k$. Let $v = \mathcal{L}^{m-1}u$. By assumption $v \in H^2_N$. By above $v = \sum_k a_k \lambda_k^{m-1} \varphi_k$.

Since φ_k is solution of $\mathcal{L}\varphi_k = \lambda_k \varphi_k$

$$\langle \mathcal{L}\varphi_k, v \rangle_\mu = \langle \lambda_k \varphi_k, v \rangle_\mu.$$

Using that $v \in H^2$, $\nabla v \cdot \vec{n} = 0$ on $\partial\Omega$ and integration by parts we obtain

$$\langle \varphi_k, \mathcal{L}v \rangle_\mu = \langle \lambda_k \varphi_k, \sum_j a_j \lambda_j^{m-1} \varphi_j \rangle_\mu = \lambda_k^m a_k.$$

ACCEPTED MANUSCRIPT

Given that $\mathcal{L}v$ is an L^2_μ function, we conclude that $\mathcal{L}v = \sum_k \lambda_k^m a_k \varphi_k$. Therefore $\sum_k \lambda_k^{2m} a_k^2 < \infty$ and hence $u \in \mathcal{H}^{2m}$.

To show the opposite inclusion, consider $u \in \mathcal{H}^{2m}$. Then $u = \sum_k a_k \varphi_k$ and $\sum_k \lambda_k^{2m} a_k^2 < \infty$. By induction step we know that $u \in H_N^{2m-2}$ and thus $v = \mathcal{L}^{m-1} u \in L^2$. We conclude as before that $v = \sum_k \lambda_k^{m-1} a_k \varphi_k$. Let $b_k = \lambda_k^{m-1} a_k$. Assumptions on u imply $\sum_k \lambda_k^2 b_k^2 < \infty$. Arguing as above in the case $s = 1$ we conclude that the series converges in H^1 and that $\nabla v = \sum_k b_k \nabla \varphi_k$. Combining this with the fact that $\mathcal{L}\varphi_k = \lambda_k \varphi_k$ in Ω for all k implies that v is a weak solution of

$$\mathcal{L}v = \sum_k \lambda_k b_k \varphi_k \quad \text{in } \Omega,$$
$$\frac{\partial v}{\partial n} = 0 \quad \text{on } \partial\Omega.$$

Since RHS of the equation is in L^2 and $\partial\Omega$ is $C^{1,1}$, by elliptic regularity [48], $v \in H^2$ and $\|v\|_{H^2}^2 \le C(\Omega, \rho) \sum_k b_k^2 \lambda_k^2$. Furthermore v satisfies the Neumann boundary condition and thus $v \in H_N^2$.

<u>Case 2°</u> Consider s odd; that is $s = 2m + 1$ for some integer $m > 0$. Assume $u \in H_N^{2m+1}$. Let $v = \mathcal{L}^m u$. Then $v \in H^1$. The result now follows analogously to the case $s = 1$. If $u \in \mathcal{H}^{2m+1}$ then, $u = \sum_k a_k \varphi_k$ with $\sum_k \lambda_k^{2m+1} a_k^2 < \infty$. By induction hypothesis, $v = \mathcal{L}^{m-1} u \in H_N^1$ and $v = \sum_k b_k \varphi_k$ where $b_k = \lambda^{m-1} a_k$. Thus $\sum_k \lambda_k b_k^2 < \infty$ and the argument proceeds as in the case $s = 1$.

Proving the equivalence of inner products is straightforward. □

We now present the proof of Lemma 2.4.

Proof of Lemma 2.4. If s is an integer the claim follows form Lemma 7.1 and Sobolev embedding theorem. Assume $s = m + \theta$ for some $\theta \in (0, 1)$. Since Ω is Lipschitz, by extension theorem of Stein (Leoni [38] 2nd edition, Theorem 13.17) there is a bounded linear extension mapping $E_m : H^m(\Omega) \to H^m(\mathbb{R}^d)$ such that $E_m(f)|_\Omega = f$. From the construction (see remark 13.9 in [38]) it follows that E_m and E_{m+1} agree on smooth functions and thus

ACCEPTED MANUSCRIPT

$E_{m+1} = E_m|_{H^m(\Omega)}$. Therefore, by Theorem 16.12 in Leoni's book (or Lemma 3.7 of Abels [32]) E_m provides a bounded mapping from the interpolation space $[H^m(\Omega), H^{m+1}(\Omega)]_{\theta,2} \to [H^m(\mathbb{R}^d), H^{m+1}(\mathbb{R}^d)]_{\theta,2}$. As discussed above the statement of Lemma 2.4 $\mathcal{H}^{m+\theta}(\Omega) = [\mathcal{H}^m(\Omega), \mathcal{H}^{m+1}(\Omega)]_{\theta,2}$. By Lemma 7.1, $[\mathcal{H}^m(\Omega), \mathcal{H}^{m+1}(\Omega)]_{\theta,2}$ embeds into $[H^m(\Omega), H^{m+1}(\Omega)]_{\theta,2}$. Furthermore, we use that, see Abels [32] Corollary 4.15, $[H^m(\mathbb{R}^d), H^{m+1}(\mathbb{R}^d)]_{\theta,2} = H^{m+\theta}(\mathbb{R}^d)$. Combining these facts yields the existence of an bounded, linear, extension mapping $\mathcal{H}^{m+\theta}(\Omega) \to H^{m+\theta}(\mathbb{R}^d)$. The results (i) and (ii) follows by the Sobolev embedding theorem. □

7.2. Passage from Discrete to Continuum

There are two key tools we use to pass from the discrete to continuum limit. The first is Γ-convergence. Γ-convergence was introduced in the 1970's by De Giorgi as a tool for studying sequences of variational problems. More recently this methodology has been applied to study the large data limits of variational problems that arise from statistical inference, e.g. [25, 49, 50, 20, 51]. Accessible introductions to Γ-convergence can be found in [41, 52]

The Γ-convergence methodology provides a notion of convergence of functionals that captures the behaviour of minimizers. In particular the minimizers converge along a subsequence to a minimizer of the limiting functional. In our setting, the objects of interest are functions on discrete domains and hence it is not immediate how one should define convergence. This brings us to our second key tool. Recently a suitable topology has been identified to characterize the convergence of discrete to continuum using an optimal transport framework [49]. The main idea is, given a discrete function $u_n : \Omega_n \to \mathbb{R}$ and a continuum function $u : \Omega \to \mathbb{R}$, to include the measures with respect to which they are defined in the comparison. Namely, one can think of the function u_n as belonging to the L^p space over the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and u belonging to the L^p space over the measure μ. One defines a continuum function $\tilde{u}_n : \Omega \to \mathbb{R}$ by $\tilde{u}_n = u_n \circ T_n$ where $T_n : \Omega_n \to \Omega$ is a measure preserving map between μ and μ_n. One then compares u_n and $\tilde{u}_n$ in the L^p distance, and simultaneously compares

ACCEPTED MANUSCRIPT

T_n and identity. In other words one considers both the difference in values and the how far the matched points are. We give a brief overview of Γ-convergence and the TL^p space.

7.2.1. A Brief Introduction to Γ-Convergence

We present the definition of Γ-convergence in terms of an abstract topology. In the next section we will discuss what topology we will use in our results. For now, we simply point out that the space $\mathcal{X}$ needs to be general enough to include functions defined with respect to different measures.

Definition 7.1. *Given a topological space $\mathcal{X}$, we say that a sequence of functions $F_n : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ Γ-converges to $F_\infty : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$, and we write $F_\infty = \Gamma\text{-}\lim_{n\to\infty} F_n$, if the following two conditions hold:*

- *(the liminf inequality) for any convergent sequence $u_n \to u$ in $\mathcal{X}$*

$$\liminf_{n\to\infty} F_n(u_n) \geq F_\infty(u);$$

- *(the limsup inequality) for every $u \in \mathcal{X}$ there exists a sequence u_n in $\mathcal{X}$ with $u_n \to u$ and*

$$\limsup_{n\to\infty} F_n(u_n) \leq F_\infty(u).$$

In the above definition we also call any sequence $\{u_n\}_{n=1,\ldots}$ that satisfies the limsup inequality a recovery sequence. The justification of Γ-convergence as the natural setting to study sequences of variational problems is given by the next proposition. The proof can be found in, for example, [41].

Proposition 7.2. *Let $F_n, F_\infty : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$. Assume that F_∞ is the Γ-limit of F_n and the sequence of minimizers $\{u_n\}_{n=1,\ldots}$ of F_n is precompact. Then*

$$\lim_{n\to\infty} \min_{\mathcal{X}} F_n = \lim_{n\to\infty} F_n(u_n) = \min_{\mathcal{X}} F_\infty$$

and furthermore, any cluster point u of $\{u_n\}_{n=1,\ldots}$ is a minimizer of F_∞.

Note that $\Gamma\text{-}\lim_{n\to\infty} F_n = F_\infty$ and $\Gamma\text{-}\lim_{n\to\infty} G_n = G_\infty$ do not imply $F_n + G_n$ Γ-converges to $G_\infty + F_\infty$. Hence, in order to build optimization problems by

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

considering individual terms it is not enough, in general, to know that each term Γ-converges. In particular, we consider using the quadratic form $J_n^{(\alpha,\tau)}$ as a prior and adding fidelity terms, e.g.

$$\mathsf{J}^{(n)}(u) = J_n^{(\alpha,\tau)}(u) + \Phi^{(n)}(u).$$

We show that, with probability one, $\Gamma\text{-}\lim_{n\to\infty} J_n^{(\alpha,\tau)} = J_\infty^{(\alpha,\tau)}$. In order to show that $\mathsf{J}^{(n)}$ Γ-converges it suffices to show that $\Phi^{(n)}$ converges along any sequence (μ_n, u_n) along which $J_n^{(\alpha,\tau)}(u_n)$ is finite. This is similar to the notion of continuous convergence, which is typically used [52, Proposition 6.20]. However we note that $\Phi^{(n)}$ does not converge continuously since as a functional on $TL^p(\Omega)$ it takes the value infinity whenever the measure considered is not μ_n.

7.2.2. The TL^p Space

In this section we give an overview of the topology that was introduced in [49] to compare sequences of functions on graphs. We motivate the topology in the setting considered in this paper. Recall that $\mu \in \mathcal{P}(\Omega)$ has density ρ and that μ_n is the empirical measure. Given $u_n : \Omega_n \to \mathbb{R}$ and $u : \Omega \to \mathbb{R}$ the idea is to consider pairs (μ, u) and (μ_n, u_n) and compare them as such. We define the metric as follows.

Definition 7.2. *Given a bounded open set Ω, the space $TL^p(\Omega)$ is the space of pairs (μ, f) such that μ is a probability measure supported on Ω and $f \in L^p(\mu)$. The metric on TL^p is defined by*

$$d_{TL^p}((f,\mu),(g,\nu)) = \inf_{\pi\in\Pi(\mu,\nu)} \left(\int_{\Omega\times\Omega} |x-y|^p + |f(x)-g(y)|^p \,\mathrm{d}\pi(x,y) \right)^{\frac{1}{p}}.$$

Above $\Pi(\mu,\nu)$ is the set of transportation plans (i.e. couplings) between μ and ν; that is the set of probability measures on $\Omega\times\Omega$ whose first marginal is μ and second marginal in ν.

For a proof that d_{TL^p} is a metric on TL^p see [49, Remark 3.4].

To connect the TL^p metric defined above with the ideas discussed previously we make several observations. The first is that when μ has a continuous density

ACCEPTED MANUSCRIPT

then one can consider transport maps $T:\Omega\to\Omega_n$ that satisfy $T_{\#}\mu=\mu_n$ instead of transport plans $\pi\in\Pi(\mu,\nu)$. Hence, one can show that

$$d_{TL^p}((f,\mu),(g,\nu)) = \inf_{T:T_{\#}\mu=\nu}\left(\|\mathrm{Id}-T\|^p_{L^p(\mu)}+\|f-g\circ T\|^p_{L^p(\mu)}\right)^{\frac{1}{p}}.$$

In the setting when we compare (μ,u) and (μ_n,u_n) the second term is nothing but $\|u-\tilde{u}_n\|^p_{L^p(\mu)}$, where $\tilde{u}_n=u_n\circ T_n$ and $T_n:\Omega\to\Omega_n$ is a transport map.

We note that for a sequence (μ_n,u_n) to TL^p converge to (μ,u) it is necessary that $\|\mathrm{Id}-T\|_{L^p(\mu)}$ converges to zero, in other words it is necessary that the measures μ_n converge to μ in p-optimal transportation distance. We recall that since Ω is bounded this is equivalent to weak convergence of μ_n to μ. Assuming this to be the case, we call any sequence of transportation maps T_n satisfying $(T_n)_{\#}\mu=\mu_n$ and $\|\mathrm{Id}-T_n\|_{L^p(\mu)}\to 0$ a stagnating sequence. One can then show (see [49, Proposition 3.12]) that convergence in TL^p is equivalent to weak* convergence of measures μ_n to μ and convergence $\|u-u_n\circ T_n\|_{L^p(\mu)}\to 0$ for arbitrary sequence of stagnating transportation maps. Furthermore if convergence $\|u-u_n\circ T_n\|_{L^p(\mu)}\to 0$ holds for a sequence of stagnating transportation maps it holds for every sequence of stagnating transportation maps.

The intrinsic scaling of the graph Laplacian, i.e. the parameter ε_n, depends on how far one needs to move "mass" to couple μ and μ_n, that is on upper bounds on transportation distance between μ and μ_n. The following result can be found in [53], the lower bound in the scaling of $\varepsilon=\varepsilon_n$ is so that there exists a stagnating sequence of transport maps with $\frac{\|T_n-\mathrm{Id}\|_{L^\infty}}{\varepsilon_n}\to 0$.

Proposition 7.3. *Let $\Omega\subset\mathbb{R}^d$ with $d\geq 2$ be open, connected and bounded with Lipschitz boundary. Let $\mu\in\mathcal{P}(\Omega)$ with density ρ which is bounded above and below by strictly positive constants. Let $\Omega_n=\{x_i\}_{i=1}^n$ where $x_i\overset{\text{iid}}{\sim}\mu$ and let $\mu_n=\frac{1}{n}\sum_{i=1}^n\delta_{x_i}$ be the associated empirical measure. Then, there exists $C>0$ such that, with probability one, there exists a sequence of transportation maps $T_n:\Omega\to\Omega_n$ that pushes μ onto μ_n and such that*

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

ACCEPTED MANUSCRIPT

where

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

7.3. Estimates on Eigenvalues of the Graph Laplacian

The following lemma is nonasymptotic and holds for all n. However we will use it in the asymptotic regime and note that our assumptions on ε, (5), and results of Proposition 7.3 ensure that the assumptions of the lemma are satisfied.

Lemma 7.4. *Consider the operator $A^{(n)}$ defined in* (1) *for $\alpha = 1$ and $\tau \geq 0$. Assume that $d_{\mathrm{OT}^\infty}(\mu_n, \mu) < \varepsilon$. Then the spectral radius λ_{max} of $A^{(n)}$ is bounded by $C\frac{1}{\varepsilon^2} + \tau^2$ where $C > 0$ is independent of n and ε.*

Let $R > 0$ be such that $\eta(3R) > 0$. Assume that $d_{\mathrm{OT}^\infty}(\mu_n, \mu) < R\varepsilon$. Then there exists $c > 0$, independent of n and ε, such that $\lambda_{max} > c\frac{1}{\varepsilon^2} + \tau^2$.

Proof. Let $\overline{\eta}(x) = \eta((|x| - 1)_+)$. Note that $\overline{\eta} \geq \eta(|\cdot|)$ and that since η is decreasing and integrable $\int_{\mathbb{R}^d} \overline{\eta}(x)dx < \infty$.

Let T be the d_{OT^∞} transport map from μ to μ_n. By assumption $\|T_n(x) - x\| \leq \varepsilon$ a.e. By definition of $A^{(n)}$

$$\lambda_{max} = \sup_{\|u\|_{L^2_{\mu_n}}=1} \langle u, A^{(n)}u\rangle_{\mu_n} = \tau^2 + \sup_{\|u\|_{L^2_{\mu_n}}=1} \langle u, s_n L u\rangle_{\mu_n}$$

We estimate

$$\begin{aligned}
\sup_{\|u\|_{L^2_{\mu_n}}=1} \langle u, s_n L u\rangle_{\mu_n} &\leq \sup_{\frac{1}{n}\sum_{i=1}^n u_i^2 = 1} \frac{4}{\sigma_\eta} \sum_{i,j} \frac{1}{n^2 \varepsilon^{d+2}} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right)(u_i^2 + u_j^2) \\
&\lesssim \sup_{\frac{1}{n}\sum_{i=1}^n u_i^2 = 1} \sum_{i=1}^n \sum_{j=1}^n \frac{1}{n^2 \varepsilon^{d+2}} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right) u_i^2 \\
&= \sup_{\frac{1}{n}\sum_{i=1}^n u_i^2 = 1} \frac{1}{n \varepsilon^{d+2}} \sum_{i=1}^n u_i^2 \int_\Omega \eta\left(\frac{|x_i - T(x)|}{\varepsilon}\right) d\mu(x) \\
&\leq \sup_{\frac{1}{n}\sum_{i=1}^n u_i^2 = 1} \frac{1}{n \varepsilon^{d+2}} \sum_{i=1}^n u_i^2 \int_\Omega \overline{\eta}\left(\frac{x_i - x}{\varepsilon}\right) d\mu(x) \\
&\lesssim \frac{1}{\varepsilon^2} \int_{\mathbb{R}^d} \overline{\eta}(z)dz \lesssim \frac{1}{\varepsilon^2}.
\end{aligned}$$

Above $\lesssim$ means $\leq$ up to a factor independent of ε and n.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

To prove the second claim of the lemma consider $v = \sqrt{n}\delta_{x_i}$, a singleton concentrated at an arbitrary x_i, that is $v_i = \sqrt{n}$ and $v_j = 0$ for all $j \neq i$. Then $\|v\|_{L^2_{\mu_n}} = 1$. Using that for a.e. $x \in B(x_i, 2\varepsilon R)$, $|x_i - T(x)| \leq 3\varepsilon R$ we estimate:

$$\begin{aligned}\sup_{\|u\|_{L^2_{\mu_n}}=1} \langle u, s_n L u\rangle_{\mu_n} &\geq \langle v, s_n L v\rangle_{\mu_n} \\ &\gtrsim \sum_{j\neq i} \frac{n}{n^2\varepsilon^{d+2}} \eta\left(\frac{|x_i - x_j|}{\varepsilon}\right) \\ &= \frac{1}{\varepsilon^{d+2}} \int_{\Omega\setminus T^{-1}(x_i)} \eta\left(\frac{|x_i - T(x)|}{\varepsilon}\right) d\mu(x) \\ &\geq \frac{1}{\varepsilon^{d+2}} \int_{B(x_i,2\varepsilon R)\setminus B(x_i,\varepsilon R)} \eta(3R)\, d\mu(x) \gtrsim \frac{1}{\varepsilon^2} \end{aligned} \tag{25}$$

which implies the claim. □

An immediate corollary of the claim is the characterization of the energy of a singleton. For any $\alpha \geq 1$ and $\tau \geq 0$.

$$J_n^{(\alpha,\tau)}(\delta_{x_i}) \sim \frac{1}{n}\left(\frac{1}{\varepsilon_n^2} + \tau^2\right)^{\alpha} \sim \frac{1}{n\varepsilon_n^{2\alpha}}. \tag{26}$$

The upper bound is immediate from the first part of the lemma, while the lower bound follows from the second part of the lemma via Jensen's inequality. Namely, $(\lambda_k^{(n)}, q_k^{(n)})$ be eigenpairs of L and let us expand δ_{x_i} in the terms of $q_k^{(n)}$: i.e. $\delta_{x_i} = \sum_{k=1}^n a_k q_k^{(n)}$ where $\sum_k a_k^2 = \|\delta_{x_i}\|^2_{L^2_{\mu_n}} = \frac{1}{n}$. We know that $\sum_k \lambda_k^{(n)} a_k^2 \gtrsim \frac{1}{n\varepsilon_n^2 s_n} \sim 1$, from (25) (using the expansion (27) and noting that $v = \sqrt{n}\delta_{x_i}$ in (25)). Hence

$$J_n^{(\alpha,\tau)}(\delta_{x_i}) = \frac{1}{2n}\sum_{k=1}^n \left(s_n\lambda_k^{(n)} + \tau^2\right)^{\alpha} n a_k^2 \geq \frac{1}{2n}\left(n s_n \sum_{k=1}^n \lambda_k^{(n)} a_k^2 + \tau^2\right)^{\alpha} \geq \frac{1}{2n}\left(\frac{1}{\varepsilon_n^2} + \tau^2\right)^{\alpha}.$$

7.4. The Limiting Quadratic Form

Here we prove Theorem 2.2. The key tool is to use spectral decomposition of the relevant quadratic forms, and to rely on the limiting properties of the eigenvalues and eigenvectors of L established in [25].

Let $(q_k^{(n)}, \lambda_k^{(n)})$ be eigenpairs of L with eigenvalues λ_k ordered so that

$$0 = \lambda_1^{(n)} \leq \lambda_2^{(n)} \leq \lambda_3^{(n)} \leq \ldots \lambda_n^{(n)}$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

where $\lambda_1^{(n)} < \lambda_2^{(n)}$ provided that the graph G is connected. We extend $F : \mathbb{R} \mapsto \mathbb{R}$ to a matrix-valued function F via $F(L) = Q^{(n)}(\Lambda_F^{(n)})(Q^{(n)})^*$ where $Q^{(n)}$ is the matrix with columns $\{q_k^{(n)}\}_{k=1}^n$ and $\Lambda_F^{(n)}$ is the diagonal matrix with entries $\{F(\lambda_i^{(n)})\}_{i=1}^n$. For constants $\alpha \geq 1$, $\tau \geq 0$ and a scaling factor s_n, given by (6), we recall the definition of the precision matrix $A^{(n)}$ is $A^{(n)} = (s_n L + \tau^2 I)^\alpha$ and the fractional Sobolev energy $J_n^{(\alpha,\tau)}$ is

$$J_n^{(\alpha,\tau)} : L^2_{\mu_n} \mapsto [0, +\infty), \qquad J_n^{(\alpha,\tau)}(u) = \frac{1}{2}\langle u, A^{(n)} u\rangle_{\mu_n}.$$

Note that

$$J_n^{(\alpha,\tau)}(u) = \frac{1}{2}\sum_{k=1}^n (s_n \lambda_k^{(n)} + \tau^2)^\alpha \langle u, q_k^{(n)}\rangle^2_{\mu_n}. \tag{27}$$

When showing Γ-convergence, all functionals are considered as functionals on the TL^p space. When evaluating $J_n^{(\alpha,\tau)}$ at (ν, u) we consider it infinite for any measure ν other than μ_n, and having the value $J_n^{(\alpha,\tau)}(u)$ defined above if $\nu = \mu_n$.

We let (q_k, λ_k) for $k = 1, 2, \ldots$ be eigenpairs of $\mathcal{L}$ ordered so that

$$0 = \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots.$$

We extend $F : \mathbb{R} \mapsto \mathbb{R}$ to an operator valued function via the identity $F(\mathcal{L}) = \sum_{k=1}^\infty F(\lambda_k)\langle u, q_k\rangle_\mu q_k$. For constants $\alpha \geq 1$ and $\tau \geq 0$ we recall the definition of the precision operator $\mathcal{A}$ as $\mathcal{A} = (\mathcal{L} + \tau I)^\alpha$ and the continuum Sobolev energy $J_\infty^{(\alpha,\tau)}$ as

$$J_\infty^{(\alpha,\tau)} : L^2_\mu \mapsto \mathbb{R} \cup \{+\infty\}, \qquad J_\infty^{(\alpha,\tau)}(u) = \frac{1}{2}\langle u, \mathcal{A}u\rangle_\mu.$$

Note that the Sobolev energy can be written

$$J_\infty^{(\alpha,\tau)}(u) = \frac{1}{2}\sum_{k=1}^\infty (\lambda_k + \tau^2)^\alpha \langle u, q_k\rangle^2_\mu.$$

Proof of Theorem 2.2. We prove the theorem in three parts. In the first part we prove the liminf inequality and in the second part the limsup inequality. The third part is devoted to the proof of the two compactness results.

The Liminf Inequality. Let $u_n \to u$ in TL^p, we wish to show that

$$\liminf_{n\to\infty} J_n^{(\alpha,\tau)}(u_n) \geq J_\infty^{(\alpha,\tau)}(u).$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

By [25, Theorem 1.2], if all eigenvalues of $\mathcal{L}$ are simple, we have with probability one (where the set of probability one can be chosen independently of the sequence u_n and u) that $s_n\lambda_k^{(n)} \to \lambda_k$ and $q_k^{(n)}$ converge in TL^2 to q_k. If there are eigenspaces of $\mathcal{L}$ of dimension higher than one then $q_k^{(n)}$ converge along a subsequence in TL^2 to eigenfunctions $\tilde{q}_k$ corresponding to the same eigenvalue as q_k. In this case we replace q_k by $\tilde{q}_k$, which does not change any of the functionals considered. We note that while eigenvectors in the general case only converge along subsequences, the projections to the relevant spaces of eigenvectors converge along the whole sequence, see [25, statement 3. Theorem 1.2]. To prove the convergence of the functional one would need to use these projections, which makes the proof cumbersome. For that reason in the remainder of the proof we assume that all eigenvalues of $\mathcal{L}$ are simple, in which case we can express the projections using the inner product with eigenfunctions.

Since $q_k^{(n)} \to q_k$ and $u_n \to u$ in TL^2 as $n \to \infty$, $\langle q_k^{(n)}, u_n\rangle_{\mu_n} \to \langle q, u\rangle_\mu$ as $n \to \infty$.

First we assume that $J_\infty^{(\alpha,\tau)}(u) < \infty$. Let $\delta > 0$ and choose K such that

$$\frac{1}{2}\sum_{k=1}^{K}(\lambda_k + \tau^2)^\alpha \langle u, q_k\rangle_\mu^2 \geq J_\infty^{(\alpha,\tau)}(u) - \delta.$$

Now,

$$\begin{aligned}\liminf_{n\to\infty} J_n^{(\alpha,\tau)}(u_n) &\geq \liminf_{n\to\infty} \frac{1}{2}\sum_{k=1}^{K}(s_n\lambda_k^{(n)} + \tau^2)^\alpha \langle u_n, q_k^{(n)}\rangle_{\mu_n}^2 \\ &= \frac{1}{2}\sum_{k=1}^{K}(\lambda_k + \tau^2)^\alpha \langle u_n, q_k\rangle_\mu^2 \\ &\geq J_\infty^{(\alpha,\tau)}(u) - \delta.\end{aligned}$$

Let $\delta \to 0$ to complete the liminf inequality for when $J_\infty^{(\alpha,\tau)}(u) < \infty$. If $J_\infty^{(\alpha,\tau)}(u) = +\infty$ then choose any $M > 0$ and find K such that $\frac{1}{2}\sum_{k=1}^{K}(\lambda_k + \tau^2)^\alpha\langle u_n, q_k\rangle_\mu^2 \geq M$, the same argument as above implies that

$$\liminf_{n\to\infty} J_n^{(\alpha,\tau)}(u_n) \geq M$$

and therefore $\liminf_{n\to\infty} J_n^{(\alpha,\tau)}(u_n) = +\infty$.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

The Limsup Inequality. As above, we assume for simplicity, that all eigenvalues of $\mathcal{L}$ are simple. We remark that there are no essential difficulties to carry out the proof in the general case.

Let $u \in L^2_\mu$ with $J_\infty^{(\alpha,\tau)}(u) < \infty$ (the proof is trivial if $J_\infty^{(\alpha,\tau)} = \infty$). Define $u_n \in L^2_{\mu_n}$ by $u_n = \sum_{k=1}^{K_n} \psi_k q_k^{(n)}$ where $\psi_k = \langle u, q_k \rangle_\mu$. Let T_n be the transport maps from μ to μ_n as in Proposition 7.3. Let $a_k^n = \psi_k q_k^{(n)} \circ T_n$ and $a_k = \psi_k q_k$. By Lemma 7.7, there exists a sequence $K_n \to \infty$ such that u_n converges to u in TL^2 metric.

We recall from the proof of the liminf inequality that $\langle q_k^{(n)}, u_n \rangle_{\mu_n} \to \langle q_k, u \rangle_\mu$ as $n \to \infty$. Combining with the convergence of eigenvalues, [25, Theorem 1.2], implies

$$(s_n \lambda_k^{(n)} + \tau^2)^\alpha \langle u_n, q_k^{(n)} \rangle_{\mu_n}^2 \to (\lambda_k + \tau^2)^\alpha \langle u, q_k \rangle_\mu^2$$

as $n \to \infty$. Taking $a_k^n = (s_n \lambda_k^{(n)} + \tau^2)^\alpha \langle u_n, q_k^{(n)} \rangle_{\mu_n}^2$ and $a_k = (\lambda_k + \tau^2)^\alpha \langle u, q_k \rangle_\mu^2$ and using Lemma 7.7 implies that there exists $\tilde{K}_n \le K_n$ converging to infinity such that $\sum_{k=1}^{\tilde{K}_n} a_k^n \to \sum_{k=1}^\infty a_k$ as $n \to \infty$. Let $\tilde{u}_n = \sum_{k=1}^{\tilde{K}_n} \psi_k q_k^{(n)}$. Then $\tilde{u}_n \to u$ in TL^2. Furthermore $J_n^{(\alpha,\tau)}(\tilde{u}_n) = \sum_{k=1}^{\tilde{K}_n} a_k^n$ and $J_\infty^{(\alpha,\tau)}(u) = \sum_{k=1}^\infty a_k$ which implies that $J_n^{(\alpha,\tau)}(\tilde{u}_n) \to J_\infty^{(\alpha,\tau)}(u)$ as $n \to \infty$.

Compactness. If $\tau > 0$ and $\sup_{n \in \mathbb{N}} J_n^{(\alpha,\tau)}(u_n) \le C$ then

$$\tau^{2\alpha} \|u_n\|^2_{L^2_{\mu_n}} = \tau^{2\alpha} \sum_{k=1}^n \langle u_n, q_k^{(n)} \rangle_{\mu_n}^2 \le \sum_{k=1}^n (s_n \lambda_k^{(n)} + \tau^2)^\alpha \langle u_n, q_k^{(n)} \rangle_{\mu_n}^2 \le C.$$

Therefore $\|u_n\|_{L^2_{\mu_n}}$ is bounded. Hence in statements 2 and 3 of the theorem we have that $\|u_n\|_{L^2_{\mu_n}}$ and $J_n^{(\alpha,\tau)}(u_n)$ are bounded. That is there exists $C > 0$ such that

$$\|u\|^2_{L^2_{\mu_n}} = \sum_{k=1}^n \langle u_n, q_k^{(n)} \rangle_{\mu_n} \le C \quad \text{and} \quad s_n^\alpha \sum_{k=1}^n (\lambda_k^{(n)})^\alpha \langle u_n, q_k^{(n)} \rangle_{\mu_n}^2 \le C. \tag{28}$$

We will show there exists $u \in L^2_\mu$ and a subsequence n_m such that u_{n_m} converges to u in TL^2.

Let $\psi_k^n = \langle u_n, q_k^{(n)} \rangle_{\mu_n}$ for all $k \le n$. Due to (28) $|\psi_k^n|$ are uniformly bounded. Therefore, by a diagonal procedure, there exists a increasing sequence $n_m \to \infty$ as $m \to \infty$ such that for every k, $\psi_k^{n_m}$ converges as $m \to \infty$. Let $\psi_k =$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

$\lim_{m\to\infty}\psi_k^{n_m}$. By Fatou's lemma, $\sum_{k=1}^{\infty}|\psi_k|^2 \leq \liminf_{m\to\infty}\sum_{k=1}^{n_m}|\psi_k^{n_m}|^2 \leq C$. Therefore $u := \sum_{k=1}^{\infty}\psi_k q_k \in L^2_\mu$. Using Lemma 7.7 and arguing as in the proof of the limsup inequality we obtain that there exists a sequence K_m increasing to infinity such that $\sum_{k=1}^{K_m}\psi_k^{n_m}q_k^{(n_m)}$ converges to u in TL^2 metric as $m\to\infty$. To show that u_{n_m} converges to u in TL^2, we now only need to show that $\|u_{n_m} - \sum_{k=1}^{K_m}\psi_k^{n_m}q_k^{(n_m)}\|_{L^2_{\mu_{n_m}}}$ converges to zero. This follows from the fact that

$$\sum_{k=K_m+1}^{n_m}|\psi_k^{n_m}|^2 \leq \frac{1}{\left(\lambda_{K_m}^{(n_m)}\right)^\alpha}\sum_{k=K_m+1}^{n_m}(\lambda_k^{(n_m)})^\alpha|\psi_k^{n_m}|^2 \leq \frac{C}{\left(s_{n_m}\lambda_{K_m}^{(n_m)}\right)^\alpha}$$

using that the sequence of eigenvalues is nondecreasing. Now since $s_{n_m}\lambda_{K_m}^{(n_m)} \geq s_{n_m}\lambda_K^{(n_m)} \to \lambda_K$ for all $K_m \geq K$, and $\lim_{K\to\infty}\lambda_K = +\infty$ we have that $s_{n_m}\lambda_{K_m}^{(n_m)} \to +\infty$ as $m\to\infty$, hence u_{n_m} converges to u in TL^2. □

Remark 7.5. *Note that when $\alpha \geq 1$ the compactness property holds trivially from the compactness property for $\alpha = 1$, see [25, Theorem 1.4], as $J_n^{(\alpha,\tau)}(u_n) \geq J_n^{(1,0)}(u_n)$.*

7.5. Variational Convergence of Probit in Labelling Model 1

To prove minimizers of the Probit model in **Labelling Model 1** converge we apply Proposition 7.2. This requires us to show that $\mathsf{J}_{\mathrm{p}}^{(n)}$ Γ-converges to $\mathsf{J}_{\mathrm{p}}^{(\infty)}$ and the compactness of sequences of minimizers. Recall that $\mathsf{J}_{\mathrm{p}}^{(n)} = J_n^{(\alpha,\tau)} + \frac{1}{n}\Phi_{\mathrm{p}}^{(n)}(\cdot;\gamma)$. Hence Theorem 2.2 establishes the Γ-convergence of the first term. We now show that $\frac{1}{n}\Phi_{\mathrm{p}}^{(n)}(u_n;y_n;\gamma) \to \Phi_{\mathrm{p},1}(u;y;\gamma)$ whenever $(\mu_n,u_n)\to(\mu,u)$ in the TL^2 sense, which is enough to establish Γ-convergence. Namely since, by definition, $J_n^{(\alpha,\tau)}$ applied to an element $(\nu,v)\in TL^p(\Omega)$ is ∞ if $\nu\neq\mu_n$ it suffices to consider sequences of the form (μ_n,u_n) to show the liminf inequality. The limsup inequality is also straightforward since the the recovery sequence for $J_\infty^{(\alpha,\tau)}$ is also of the form (μ_n,u_n).

Lemma 7.6. *Consider domain Ω and measure μ satisfying Assumptions 2–3. Let $x_i \overset{\text{iid}}{\sim} \mu$ for $i = 1,\dots,n$, $\Omega_n = \{x_1,\dots,x_n\}$ and μ_n be the empirical measure of the sample. Let Ω' be an open subset of Ω, $\mu_n' = \mu_n\lfloor_{\Omega'}$ and $\mu' = \mu\lfloor_{\Omega}$. Let*

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

$y_n \in L^\infty(\mu_n')$ and $y \in L^\infty(\mu')$ and let $\hat{y}_n \in L^\infty(\mu_n)$ and $\hat{y} \in L^\infty(\mu)$ be their extensions by zero. Assume

$$(\mu_n, \hat{y}_n) \to (\mu, \hat{y}) \quad \text{in } TL^\infty \text{ as } n \to \infty.$$

Let $\Phi_{\mathrm{p}}^{(n)}$ and $\Phi_{\mathrm{p},1}$ be defined by (9) and (16) respectively, where $Z' = \{j : x_j \in \Omega'\}$ and $\gamma > 0$ (and where we explicitly include the dependence of y_n and y in $\Phi_{\mathrm{p}}^{(n)}$ and $\Phi_{\mathrm{p},1}$).

Then, with probability one, if $(\mu_n, u_n) \to (\mu, u)$ in TL^p then

$$\frac{1}{n}\Phi_{\mathrm{p}}^{(n)}(u_n; y_n; \gamma) \to \Phi_{\mathrm{p},1}(u; y; \gamma) \quad \text{as } n \to \infty.$$

Proof. Let $(\mu_n, u_n) \to (\mu, u)$ in TL^p. We first note that since $\Psi(uy;\gamma) = \Psi\left(\frac{uy}{\gamma}; 1\right)$ and since multiplying all functions by a constant does not affect the TL^p convergence, it suffices to consider $\gamma = 1$. For brevity, we omit γ in the functionals that follow. We have that $\hat{y}_n \circ T_n \to \hat{y}$ and $u_n \circ T_n \to u$. Recall that

$$\begin{aligned} \frac{1}{n}\Phi_{\mathrm{p}}^{(n)}(u_n; y_n) &= \int_{T_n^{-1}(\Omega_n')} \log \Psi(y_n(T_n(x)) u_n(T_n(x)))\, \mathrm{d}\mu(x) \\ \Phi_{\mathrm{p},1}(u; y) &= \int_{\Omega'} \log \Psi(y(x)u(x))\, \mathrm{d}\mu(x), \end{aligned}$$

where $\Omega_n' = \{x_i : x_i \in \Omega',\ \text{for } i = 1, \dots, n\}$. Recall also that symmetric difference of sets is denoted by $A \triangle B = (A \setminus B) \cup (B \setminus A)$. It follows that

$$\begin{aligned} \left| \frac{1}{n}\Phi_{\mathrm{p}}^{(n)}(u_n; y_n) - \Phi_{\mathrm{p},1}(u; y) \right| &\le \left| \int_{\Omega' \triangle T_n^{-1}(\Omega_n')} \log \Psi(\hat{y}(x)u(x)) \mathrm{d}\mu(x) \right| \\ &+ \left| \int_{T_n^{-1}(\Omega_n')} \log\left(\Psi(y_n(T_n(x)) u_n(T_n(x)); \gamma\right) - \log\left(\hat{y}(x)u(x)\right) \mathrm{d}\mu(x) \right|. \end{aligned} \tag{29}$$

Define

$$\partial_{\varepsilon_n}\Omega' = \{x : \mathrm{dist}(x, \partial\Omega') \le \varepsilon_n\}.$$

Then $\Omega' \triangle T_n^{-1}(\Omega_n') \subseteq \partial_{\varepsilon_n}\Omega'$. Since $\hat{y} \in L^\infty$ and $u \in L^2_\mu$ then $\hat{y}u \in L^2_\mu$ and so by Corollary 7.9 $\log \Psi(\hat{y}u) \in L^1$. Hence, by the dominated convergence theorem

$$\left| \int_{\Omega' \triangle T_n^{-1}(\Omega_n')} \log \Psi(\hat{y}(x)u(x)) \mathrm{d}\mu(x) \right| \le \int_{\partial_{\varepsilon_n}\Omega'} |\log \Psi(\hat{y}(x)u(x))|\, \mathrm{d}\mu(x) \to 0.$$

ACCEPTED MANUSCRIPT

We are left to show that the second term on the right hand side of (29) converges to 0. Let $F(w,v) = |\log\Psi(w) - \log\Psi(v)|$. Let $M \geq 1$ and define the following sets

$$
\begin{aligned}
\mathcal{A}_{n,M} &= \left\{x \in T_n^{-1}(\Omega_n') \,:\, \min\{\hat{y}(x)u(x), y_n(T_n(x))u_n(T_n(x))\} \geq -M\right\}\\
\mathcal{B}_{n,M} &= \left\{x \in T_n^{-1}(\Omega_n') \,:\, \hat{y}(x)u(x) \geq y_n(T_n(x))u_n(T_n(x)) \leq -M\right\}\\
\mathcal{C}_{n,M} &= \left\{x \in T_n^{-1}(\Omega_n') \,:\, y_n(T_n(x))u_n(T_n(x)) \geq \hat{y}(x)u(x) \leq -M\right\}.
\end{aligned}
$$

The quantity we want to estimate satisfies

$$
\begin{aligned}
&\left|\int_{T_n^{-1}(\Omega_n')} \log\left(\Psi(y_n(T_n(x))u_n(T_n(x)))\right) - \log\Psi\left(\hat{y}(x)u(x)\right)\,\mathrm{d}\mu(x)\right|\\
&\qquad\qquad \leq \int_{T_n^{-1}(\Omega_n')} F(y_n(T_n(x))u_n(T_n(x)), \hat{y}(x)u(x))\,\mathrm{d}\mu(x).
\end{aligned}
$$

Since $T_n^{-1}(\Omega_n') = \mathcal{A}_{n,M} \cup \mathcal{B}_{n,M} \cup \mathcal{C}_{n,M}$ we proceed by estimating the integral over each of the sets, utilizing the bounds in Lemma 7.8.

$$
\begin{aligned}
&\int_{\mathcal{A}_{n,M}} F(y_n(T_n(x))u_n(T_n(x)), \hat{y}(x)u(x))\,\mathrm{d}\mu(x)\\
&\quad \leq \frac{1}{\int_{-\infty}^{-M} e^{-\frac{t^2}{2}}\,\mathrm{d}t} \int_{\mathcal{A}_{n,M}} |y_n(T_n(x))u_n(T_n(x)) - \hat{y}(x)u(x)|\,\mathrm{d}\mu(x)\\
&\quad \leq \frac{1}{\int_{-\infty}^{-M} e^{-\frac{t^2}{2}}\,\mathrm{d}t} \left(\|y_n\|_{L^2_{\mu_n}}\|u_n \circ T_n - u\|_{L^2_\mu} + \|u\|_{L^2_\mu}\|\hat{y}_n \circ T_n - \hat{y}\|_{L^2_\mu}\right).
\end{aligned}
$$

$$
\begin{aligned}
&\int_{\mathcal{B}_{n,M}} F(y_n(T_n(x))u_n(T_n(x)), \hat{y}(x)u(x))\,\mathrm{d}\mu(x)\\
&\quad \leq \int_{\mathcal{B}_{n,M}} 2|y_n(T_n(x))|^2|u_n(T_n(x))|^2\,\mathrm{d}\mu(x) + \frac{1}{M^2}\\
&\quad \leq 2\|\hat{y}_n\|^2_{L^\infty_{\mu_n}} \int_{\mathcal{B}_{n,M}} |u_n(T_n(x))|^2\,\mathrm{d}\mu(x) + \frac{1}{M^2}\\
&\quad \leq 4\|\hat{y}_n\|^2_{L^\infty_{\mu_n}} \left(\|u_n \circ T_n - u\|^2_{L^2_\mu} + \int_\Omega |u(x)|^2 \mathbb{I}_{|y_n(T_n(x))u_n(T_n(x))| \geq M}\,\mathrm{d}\mu(x)\right) + \frac{1}{M^2}.
\end{aligned}
$$

$$
\begin{aligned}
&\int_{\mathcal{C}_{n,M}} F(y_n(T_n(x))u_n(T_n(x)), \hat{y}(x)u(x))\,\mathrm{d}\mu(x)\\
&\quad \leq \int_{\mathcal{C}_{n,M}} 2|\hat{y}(x)|^2|u(x)|^2\,\mathrm{d}\mu(x) + \frac{1}{M^2}\\
&\quad \leq 2\|\hat{y}\|^2_{L^\infty_\mu} \int_\Omega |u(x)|^2 \mathbb{I}_{|y(x)u(x)| \geq M}\,\mathrm{d}\mu(x) + \frac{1}{M^2}.
\end{aligned}
$$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

For every subsequence there exists a further subsequence such that $(y_n \circ T_n)(u_n \circ T_n) \to yu$ pointwise a.e., hence by the dominated convergence theorem

$$\int_\Omega |u(x)|^2 \mathbb{I}_{|y_n(T_n(x))u_n(T_n(x))| \geq M} \,\mathrm{d}\mu(x) \to \int_\Omega |u(x)|^2 \mathbb{I}_{|y(x)u(x)| \geq M} \,\mathrm{d}\mu(x) \quad \text{as } n \to \infty.$$

Hence, for $M \geq 1$ fixed we have

$$\limsup_{n\to\infty} \left| \int_{T_n^{-1}(\Omega_n')} \log\left(\Psi(y_n(T_n(x))u_n(T_n(x)); \gamma) - \log\left(\hat{y}(x)u(x); \gamma\right) \mathrm{d}\mu(x) \right| \leq \frac{2}{M^2} + 6\|\hat{y}\|_{L^\infty_\mu} \int_\Omega |u(x)|^2 \mathbb{I}_{|\hat{y}(x)u(x)| \geq M} \,\mathrm{d}\mu(x).$$

Taking $M \to \infty$ completes the proof. □

The proof of Theorem 4.2 is now just a special case of the above lemma and an easy compactness result that follows from Theorem 2.2.

Proof of Theorem 4.2. The following statements all hold with probability one. Let

$$y(x) = \begin{cases} 1 & \text{if } x \in \Omega^+ \\ -1 & \text{if } x \in \Omega^-. \end{cases}$$

Since $\mathrm{dist}(\Omega^+, \Omega^-) > 0$ there exists a minimal Lipschitz extension $\hat{y} \in L^\infty$ of y to Ω. Let $y_n = y\lfloor_{\Omega_n}$ and $\hat{y}_n = \hat{y}\lfloor_{\Omega_n}$. Since

$$\begin{aligned} \|\hat{y}_n \circ T_n - \hat{y}\|_{L^\infty(\mu)} &= \mu\text{-}\operatorname*{ess\,sup}_{x\in\Omega} |\hat{y}_n(T_n(x)) - \hat{y}(x)| \\ &= \mu\text{-}\operatorname*{ess\,sup}_{x\in\Omega} |\hat{y}(T_n(x)) - \hat{y}(x)| \\ &\leq \mathrm{Lip}(\hat{y}) \|T_n - \mathrm{Id}\|_{L^\infty} \end{aligned}$$

we conclude that $(\mu_n, \hat{y}_n) \to (\mu, \hat{y})$ in TL^∞. Hence, by Lemma 7.6, $\frac{1}{n}\Phi_{\mathrm{p}}^{(n)}(u_n; \gamma) \to \Phi_{\mathrm{p},1}(u; \gamma)$ whenever $(\mu_n, u_n) \to (\mu, u)$ in TL^p. Combining with Theorem 2.2 implies that $\mathsf{J}_{\mathrm{p}}^{(n)}$ Γ-converges to $\mathsf{J}_{\mathrm{p}}^{(\infty)}$ via a straightforward argument.

If $\tau > 0$ then the compactness of minimizers follows from Theorem 2.2 using that $\sup_{n\in\mathbb{N}} \min_{v_n \in L^2_{\mu_n}} \mathsf{J}_{\mathrm{p}}^{(n)}(v_n) \leq \sup_{n\in\mathbb{N}} \mathsf{J}_{\mathrm{p}}^{(n)}(0) = \frac{1}{2}$.

When $\tau = 0$ we consider the sequence $w_n = v_n - \bar{v}_n$ where v_n is a minimizer of $\mathsf{J}_{\mathrm{p}}^{(n)}$ and $\bar{v}_n = \langle v_n, q_1 \rangle_{\mu_n} = \int_\Omega v_n(x) \,\mathrm{d}\mu_n(x)$. Then, $J_n^{(\alpha,0)}(w_n) = J_n^{(\alpha,0)}(v_n)$ and

$$\|w_n\|^2_{L^2_{\mu_n}} = \|v_n - \bar{v}_n\|^2_{L^2_{\mu_n}} = \sum_{k=2}^n \langle v_n, q_k \rangle^2_{\mu_n} \leq \frac{1}{(s_n \lambda_2^{(n)})^\alpha} J_n^{(\alpha,0)}(v_n).$$

ACCEPTED MANUSCRIPT

As in the case $\tau > 0$ the quadratic form is bounded, i.e. $\sup_{n\in\mathbb{N}} \mathsf{J}_{\mathrm{p}}^{(n)}(v_n) \le \frac{1}{2}$. Hence $J_n^{(\alpha,\tau)}(w_n) \le \frac{1}{2}$ and $\|w_n\|^2_{L^2_{\mu_n}} \le \frac{1}{\lambda_2^{\alpha}}$ for n large enough. By Theorem 2.2 w_n is precompact in TL^2. Therefore $\sup_{n\in\mathbb{N}} \|v_n\|_{L^2_{\mu_n}} \le M + \sup_{n\in\mathbb{N}} |\bar{v}_n|$ for some $M > 0$. Since $J_n^{(\alpha,\tau)}$ is insensitive to the addition of a constant, and $-1 \le y \le 1$, then for any minimiser v_n one must have $\bar{v}_n \in [-1,1]$. Hence $\sup_{n\in\mathbb{N}} \|v_n\|_{L^2_{\mu_n}} \le M + 1$ so by Theorem 2.2 $\{v_n\}$ is precompact in TL^2.

Since the minimizers of $\mathsf{J}_{\mathrm{p}}^{(\infty)}$ are unique (due to convexity, see Lemma 4.1), by Proposition 7.2 we have that the sequence of minimizers v_n of $\mathsf{J}_{\mathrm{p}}^{(n)}$ converges to the minimizer of $\mathsf{J}_{\mathrm{p}}^{(\infty)}$. □

7.6. *Variational Convergence of Probit in Labelling Model 2*

Proof of Theorem 4.3. It suffices to show that $\mathsf{J}_{\mathrm{p}}^{(n)}$ Γ-converges in TL^2 to $J_\infty^{(\alpha,\tau)}$ and that the sequence of minimizers v_n of $\mathsf{J}_{\mathrm{p}}^{(n)}$ is precompact in TL^2. We note that the liminf statement of the Γ-convergence follows immediately from statement 1. of Theorem 2.2.

To complete the proof of Γ-convergence it suffices to construct a recovery sequence. The strategy is analogous to the one of the proof on Theorem 4.9 of [39]. Let $v \in \mathcal{H}^\alpha(\Omega)$. Since $J_n^{(\alpha,\tau)}$ Γ-converges to $J_\infty^{(\alpha,\tau)}$ by Theorem 2.2 there exists Let $v^{(n)} \in L^2_{\mu_n}$ such that $J_n^{(\alpha,\tau)}(v^{(n)}) \to J_\infty^{(\alpha,\tau)}(v)$ as $n \to \infty$. Consider the functions

$$\tilde{v}^{(n)}(x_i) = \begin{cases} c_n y(x_i) & \text{if } i = 1,\dots,N. \\ v^{(n)}(x_i) & \text{if } i = N+1,\dots,n \end{cases}$$

where $c_n \to \infty$ and $\frac{c_n}{\varepsilon_n^{2\alpha} n} \to 0$ as $n \to \infty$.

Note that condition (5) implies that when $\alpha < \frac{d}{2}$ then (20) still holds. Therefore (26) implies that $J_n^{(\alpha,\tau)}(c_n \delta_{x_i}) \to 0$ as $n \to \infty$. Also note that since $c_n \to \infty$, $\Phi_{\mathrm{p}}^{(n)}(\tilde{v}^{(n)}; \gamma) \to 0$ as $n \to \infty$. It is now straightforward to show, using the form of the functional, the estimate on the energy of a singleton and the fact that $\varepsilon_n n^{\frac{1}{2\alpha}} \to \infty$ as $n \to \infty$, that $\mathsf{J}_{\mathrm{p}}^{(n)}(\tilde{v}^{(n)}) \to J_\infty^{(\alpha,\tau)}(v)$ as desired.

The precompactness of $\{v_n\}_{n\in\mathbb{N}}$ follows from Theorem 2.2. Since 0 is the unique minimizer of $J_\infty^{(\alpha,\tau)}$, due to $\tau > 0$, the above results imply that $v^{(n)}$ converge to 0. □

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

7.7. Small Noise Limits

Proof of Theorem 4.6. First observe that since Assumptions 2–3 hold and $\alpha > d/2$, the measure ν_0, and hence the measures $\nu_{\mathrm{p},1}, \nu_{\mathrm{p},2}, \nu_1$, are all well-defined measures on $L^2(\Omega)$ by Theorem 2.5.

(i) For any continuous bounded function $g : C(\Omega;\mathbb{R}) \to \mathbb{R}$ we have

$$\mathbb{E}^{\nu_{\mathrm{p},1}} g(u) = \frac{\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},1}(u;\gamma)} g(u)}{\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},1}(u;\gamma)}}, \quad \mathbb{E}^{\nu_1} g(u) = \frac{\mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,1}}(u) g(u)}{\mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,1}}(u)}.$$

For the first convergence it thus suffices to prove that, as $\gamma \to 0$,

$$\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},1}(u;\gamma)} g(u) \to \mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,1}}(u) g(u)$$

for all continuous functions $g : C(\Omega;\mathbb{R}) \to [-1,1]$.

We first define the standard normal cumulative distribution function $\varphi(z) = \Psi(z,1)$, and note that we may write

$$\Phi_{\mathrm{p},1}(u;\gamma) = -\int_{x\in\Omega'} \log\Big(\varphi(y(x)u(x)/\gamma)\Big)\mathrm{d}x \geq 0.$$

In what follows it will be helpful to recall the following standard Mills ratio bound: for all $t > 0$,

$$\varphi(t) \geq 1 - \frac{e^{-t^2/2}}{t\sqrt{2\pi}}. \tag{30}$$

Suppose first that $u \in B_{\infty,1}$, then $y(x)u(x)/\gamma > 0$ for a.e. $x \in \Omega'$. The assumption that $\overline{\Omega^+} \cap \overline{\Omega^-} = \varnothing$ ensures that y is continuous on $\Omega' = \Omega^+ \cup \Omega^-$. As u is also continuous on Ω', given any $\varepsilon > 0$, we may find $\Omega'_\varepsilon \subseteq \Omega'$ such that $y(x)u(x)/\gamma > \varepsilon/\gamma$ for all $x \in \Omega'_\varepsilon$. Moreover, these sets may be chosen such that $\mathrm{leb}(\Omega' \setminus \Omega'_\varepsilon) \to 0$ as $\varepsilon \to 0$. Applying the bound (30), we see that for any $x \in \Omega'_\varepsilon$,

$$\varphi(y(x)u(x)/\gamma) \geq 1 - \gamma\frac{e^{-u(x)^2 y(x)^2/2\gamma^2}}{u(x)y(x)\sqrt{2\pi}} \geq 1 - \gamma\frac{e^{-\varepsilon^2/2\gamma^2}}{\varepsilon\sqrt{2\pi}}.$$

Additionally, for any $x \in \Omega' \setminus \Omega'_\varepsilon$, we have $\varphi(y(x)u(x)/\gamma) \geq \varphi(0) = 1/2$. We deduce that

$$\begin{aligned}\Phi_{\mathrm{p},1}(u;\gamma) &= -\int_{\Omega'_\varepsilon} \log(\varphi(y(x)u(x)/\gamma)\,\mathrm{d}\mu(x) - \int_{\Omega'\setminus\Omega'_\varepsilon} \log(\varphi(y(x)u(x)/\gamma)\,\mathrm{d}\mu(x)\\ &\leq -\log\left(1 - \gamma\frac{e^{-\varepsilon^2/2\gamma^2}}{\varepsilon\sqrt{2\pi}}\right)\cdot\rho^+\cdot\mathrm{leb}(\Omega'_\varepsilon) + \log(2)\cdot\rho^+\cdot\mathrm{leb}(\Omega'\setminus\Omega'_\varepsilon).\end{aligned}$$

ACCEPTED MANUSCRIPT

The right-hand term may be made arbitrarily small by choosing ε small enough. For any given $\varepsilon > 0$, the left-hand term tends to zero as $\gamma \to 0$, and so we deduce that $\Phi_{\mathrm{p},1}(u;\gamma) \to 0$ and hence

$$e^{-\Phi_{\mathrm{p},1}(u;\gamma)} g(u) \to g(u) = \mathbb{1}_{B_{\infty,1}}(u) g(u).$$

Now suppose that $u \notin B_{\infty,1}$, and assume first that there is a subset $E \subseteq \Omega'$ with $\mathrm{leb}(E) > 0$ and $y(x)u(x) < 0$ for all $x \in E$. Then similarly to above, there exists $\varepsilon > 0$ and $E_\varepsilon \subseteq E$ with $\mathrm{leb}(E_\varepsilon) > 0$ such that $y(x)u(x)/\gamma < -\varepsilon/\gamma$ for all $x \in E_\varepsilon$. Observing that $\varphi(t) = 1 - \varphi(-t)$, we may apply the bound (30) to deduce that, for any $x \in E_\varepsilon$,

$$\varphi(y(x)u(x)/\gamma) \le -\gamma \frac{e^{-u(x)^2 y(x)^2/2\gamma^2}}{u(x)y(x)\sqrt{2\pi}} \le \frac{\gamma}{\varepsilon\sqrt{2\pi}}.$$

We therefore deduce that

$$\begin{aligned}\Phi_{\mathrm{p},1}(u;\gamma) &\ge \int_{E_\varepsilon} -\log(\varphi(y(x)u(x)/\gamma)\,\mathrm{d}\mu(x)\\ &\ge -\log\left(\frac{\gamma}{\varepsilon\sqrt{2\pi}}\right)\cdot \rho^- \cdot \mathrm{leb}(E_\varepsilon) \to \infty\end{aligned}$$

from which we see that

$$e^{-\Phi_{\mathrm{p},1}(u;\gamma)} g(u) \to 0 = \mathbb{1}_{B_{\infty,1}}(u) g(u).$$

Assume now that $y(x)u(x) \ge 0$ for a.e. $x \in \Omega'$. Since $u \notin B_{\infty,1}$ there is a subset $\Omega'' \subseteq \Omega'$ such that $y(x)u(x) = 0$ for all $x \in \Omega''$, $y(x)u(x) > 0$ a.e. $x \in \Omega' \setminus \Omega''$, and $\mathrm{leb}(\Omega'') > 0$. We then have

$$\begin{aligned}\Phi_{\mathrm{p},1}(u;\gamma) &= -\int_{\Omega''} \log(\varphi(0))\,\mathrm{d}\mu(x) - \int_{\Omega'\setminus\Omega''} \log(\varphi(y(x)u(x)/\gamma)\,\mathrm{d}\mu(x)\\ &= \log(2)\mu(\Omega'') - \int_{\Omega'\setminus\Omega''} \log(\varphi(y(x)u(x)/\gamma)\,\mathrm{d}\mu(x)\\ &\to \log(2)\mu(\Omega'').\end{aligned}$$

We hence have $e^{-\Phi_{\mathrm{p}}(u;y,\gamma)} g(u) \not\to 0 = \mathbb{1}_{B_{\infty,1}}(u) g(u)$. However, the event

$$\begin{aligned}D &:= \{u \in C(\Omega;\mathbb{R}) \,|\, \text{There exists } \Omega'' \subseteq \Omega' \text{ with } \mathrm{leb}(\Omega'') > 0 \text{ and } u|_{\Omega''} = 0\}\\ &\subseteq \{u \in C(\Omega;\mathbb{R}) \,|\, \mathrm{leb}\big(u^{-1}\{0\}\big) > 0\} = D'\end{aligned}$$

ACCEPTED MANUSCRIPT

has probability zero under ν_0. This can be deduced from Proposition 7.2 in [23]: since Assumptions 2–3 hold and $\alpha > d$, Theorem 2.5 tells us that draws from ν_0 are almost-surely continuous, which is sufficient in order to deduce the conclusions of the proposition, and so $\nu_0(D) \le \nu_0(D') = 0$. We thus have pointwise convergence of the integrand on D^c, and so using the boundedness of the integrand by 1 and the dominated convergence theorem,

$$\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},1}(u;\gamma)} g(u) = \mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},1}(u;\gamma)} g(u) \mathbb{1}_{D^c}(u) \to \mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,1}}(u) g(u)$$

which proves that $\nu_{\mathrm{p},1} \Rightarrow \nu_1$.

For the convergence $\nu_{\mathrm{ls},1} \Rightarrow \nu_1$ it similarly suffices to prove that, as $\gamma \to 0$,

$$\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{ls},1}(u;\gamma)} g(u) \to \mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,1}}(u) g(u)$$

for all continuous functions $g : C(\Omega;\mathbb{R}) \to [-1,1]$. For fixed $u \in B_{\infty,1}$ we have $e^{-\Phi_{\mathrm{ls},1}(u;\gamma)} = \mathbb{1}_{B_{\infty,1}}(u) = 1$ and hence $e^{-\Phi_{\mathrm{ls},1}(u;\gamma)} g(u) = \mathbb{1}_{B_{\infty,1}}(u) g(u)$ for all $\gamma > 0$. For fixed $u \notin B_{\infty,1}$ there is a set $E \subseteq \Omega'$ with positive Lebesgue measure on which $y(x)u(x) \le 0$. As a consequence $\Phi_{\mathrm{ls},1}(u;\gamma) \ge \frac{1}{2\gamma^2}\mathrm{leb}(E)\rho^-$ and so $e^{-\Phi_{\mathrm{ls},1}(u;\gamma)} g(u) \to 0 = \mathbb{1}_{B_{\infty,1}}(u) g(u)$ as $\gamma \to 0$. Pointwise convergence of the integrand, combined with boundedness by 1 of the integrand, gives the result.

(ii) The structure of the proof is similar to part (i). To prove $\nu_{\mathrm{p},2} \Rightarrow \nu_2$, it suffices to show that, as $\gamma \to 0$,

$$\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},2}(u;\gamma)} g(u) \to \mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,2}}(u) g(u)$$

for all continuous functions $g : C(\Omega;\mathbb{R}) \mapsto [-1,1]$. We write

$$\Phi_{\mathrm{p}}^{(n)}(u;\gamma) = -\frac{1}{n} \sum_{j \in Z'} \log\Big(\varphi(y(x_j)u(x_j)/\gamma)\Big) \ge 0.$$

Note that $\Phi_{\mathrm{p}}^{(n)}(u;\gamma)$ is well-defined almost-surely on samples from ν_0 since ν_0 is supported on continuous functions (Theorem 2.5). Suppose first that $u \in B_{\infty,2}$, then $y(x_j)u(x_j)/\gamma > 0$ for all $j \in Z'$ and $\gamma > 0$. It follows that

ACCEPTED MANUSCRIPT

for each $j \in Z'$, $y(x_j)y(x_j)/\gamma \to \infty$ as $\gamma \to 0$ and so $\varphi(y(x_j)u(x_j)/\gamma) \to 1$. Thus, $\Phi_{\mathrm{p},2}(u;\gamma) \to 0$ and so

$$e^{-\Phi_{\mathrm{p},2}(u;\gamma)} g(u) \to g(u) = \mathbb{1}_{B_{\infty,2}}(u) g(u).$$

Now suppose that $u \notin B_{\infty,2}$. Assume first that there is a $j \in Z'$ such that $y(x_j)u(x_j) < 0$, so that $y(x_j)u(x_j)/\gamma \to -\infty$ and hence $\varphi(y(x_j)u(x_j)/\gamma) \to 0$. Then we may bound

$$\Phi_{\mathrm{p},2}(u;\gamma) \geq -\log(\varphi(y(x_j)u(x_j)/\gamma) \to \infty$$

from which we see that

$$e^{-\Phi_{\mathrm{p},2}(u;\gamma)} g(u) \to 0 = \mathbb{1}_{B_{\infty,2}}(u) g(u).$$

Assume now that $y(x_j)u(x_j) \geq 0$ for all $j \in Z'$, then since $u \notin B_{\infty,2}$ there is a subcollection $Z'' \subseteq Z'$ such that $y(x_j)u(x_j) = 0$ for all $j \in Z''$ and $y(x_j)u(x_j) > 0$ for all $j \in Z' \setminus Z''$. We then have

$$\begin{aligned}
\Phi_{\mathrm{p},2}(u;\gamma) &= -\frac{1}{n} \sum_{j \in Z''} \log\Big(\varphi(0)\Big) - \frac{1}{n} \sum_{j \in Z' \setminus Z''} \log\Big(\varphi(y(x_j)u(x_j)/\gamma)\Big) \\
&= \frac{|Z''|}{n} \log(2) - \frac{1}{n} \sum_{j \in Z' \setminus Z''} \log\Big(\varphi(y(x_j)u(x_j)/\gamma)\Big) \\
&\to \frac{|Z''|}{n} \log(2).
\end{aligned}$$

Thus, in this case $e^{-\Phi_{\mathrm{p},2}(u;\gamma)} g(u) \not\to 0 = \mathbb{1}_{B_{\infty,2}}(u) g(u)$. However, the event

$$D = \{u \in C(\Omega;\mathbb{R}) \,|\, u(x_j) = 0 \text{ for some } j \in Z'\}$$

has probability zero under ν_0. To see this, observe that ν_0 is a non-degenerate Gaussian measure on $C(\Omega;\mathbb{R})$ as a consequence of Theorem 2.5. Thus $u \sim \nu_0$ implies that the vector $(u(x_1), \dots, u(x_{n^+ + n^-}))$ is a non-degenerate Gaussian random variable on $\mathbb{R}^{n^+ + n^-}$. Its law is hence equivalent to the Lebesgue measure, and so the probability that it takes value in any given hyperplane is zero. We therefore have pointwise convergence of the integrand on D^c. Since the integrand is bounded by 1, we deduce

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

from the dominated convergence theorem that

$$\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},2}(u;\gamma)} g(u) = \mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{p},2}(u;\gamma)} g(u) \mathbb{1}_{D^c}(u) \to \mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,2}}(u) g(u)$$

which proves that $\nu_{\mathrm{p},2} \Rightarrow \nu_2$.

To prove $\nu_{\mathrm{ls},2} \Rightarrow \nu_2$ we show that, as $\gamma \to 0$,

$$\mathbb{E}^{\nu_0} e^{-\Phi_{\mathrm{ls},2}(u;\gamma)} g(u) \to \mathbb{E}^{\nu_0} \mathbb{1}_{B_{\infty,2}}(u) g(u)$$

for all continuous functions $g : C(\Omega;\mathbb{R}) \mapsto [-1,1]$. For fixed $u \in B_{\infty,2}$ we have $e^{-\Phi_{\mathrm{ls},2}(u;\gamma)} = \mathbb{1}_{B_{\infty,2}}(u) = 1$ and hence $e^{-\Phi_{\mathrm{ls},2}(u;\gamma)} g(u) = \mathbb{1}_{B_{\infty,2}}(u) g(u)$ for all $\gamma > 0$. For fixed $u \notin B_{\infty,2}$ there is at least one $j \in Z'$ such that $y(x_j)u(x_j) \le 0$. As a consequence $\Phi_{\mathrm{ls},2}(u;\gamma) \ge \frac{1}{2\gamma^2}\frac{1}{n}\rho^-$ and so $e^{-\Phi_{\mathrm{ls},2}(u;\gamma)} g(u) \to 0 = \mathbb{1}_{B_{\infty,2}}(u) g(u)$ as $\gamma \to 0$. Pointwise convergence of the integrand, combined with boundedness by 1 of the integrand, gives the desired result.

□

7.8. Technical lemmas

We include technical lemmas which are used in the main Γ-convergence result (Theorem 2.2) and in the proof of convergence for the probit model.

Lemma 7.7. *Let X be a normed space and $a_k^{(n)} \in X$ for all $n \in \mathbb{N}$ and $k = 1,\dots,n$. Assume $a_k \in X$ be such that $\sum_{k=1}^{\infty} \|a_k\| < \infty$ and that for all k*

$$a_k^{(n)} \to a_k \qquad \text{as } n \to \infty.$$

Then there exists a sequence $\{K_n\}_{n=1,\dots}$ converging to infinity as $n \to \infty$ such that

$$\sum_{k=1}^{K_n} a_k^{(n)} \to \sum_{k=1}^{\infty} a_k \qquad \text{as } n \to \infty.$$

Note that if the conclusion holds for one sequence K_n it also holds for any other sequence converging to infinity and majorized by K_n.

Proof. Note that by our assumption for any fixed s, $\sum_{k=1}^{s} a_k^n \to \sum_{k=1}^{s} a_k$ as $n \to \infty$. Let K_n be the largest number such that for all $m \ge n$, $\left\| \sum_{k=1}^{K_n} a_k^{(m)} - \sum_{k=1}^{K_n} a_k \right\| <$

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

$\frac{1}{n}$. Due to observation above, $K_n \to \infty$ as $n \to \infty$. Furthermore

$$\left\| \sum_{k=1}^{K_n} a_k^n - \sum_{k=1}^{\infty} a_k \right\| \leq \left\| \sum_{k=1}^{K_n} a_k^n - \sum_{k=1}^{K_n} a_k \right\| + \left\| \sum_{k=K_n+1}^{\infty} a_k \right\|$$

which converges to zero an $n \to \infty$. □

The second result is an estimate on the behavior of the function Ψ defined in (8)

Lemma 7.8. *Let $F(w,v) = \log \Psi(w;1) - \log \Psi(v;1)$ where Ψ is defined by* (8) *with $\gamma = 1$. For all $w > v$ and $M \geq 1$,*

$$F(w,v) \leq \begin{cases} 2v^2 + \frac{1}{M^2} & \text{if } v \leq -M \\ \frac{|w-v|}{\int_{-\infty}^{-M} e^{-\frac{t^2}{2}} \, \mathrm{d}t} & \text{if } v \geq -M. \end{cases}$$

Proof. We consider the two cases: $v \leq -M$ and $v \geq -M$ separately. From inequality 7.1.13 in [54] directly follows that

$$\forall u \leq 0, \qquad \sqrt{\frac{2}{\pi}} \frac{1}{-u + \sqrt{u^2+4}} e^{-\frac{u^2}{2}} \leq \Psi(u)$$

When $v \leq -M$, by taking the logarithm we obtain

$$\begin{aligned} F(w,v) &\leq -\log \Psi(v;\gamma) \leq -\log\left(\sqrt{\frac{2}{\pi}} \frac{1}{-v + \sqrt{v^2+4}} e^{-\frac{v^2}{2}} \right) \leq \sqrt{\frac{\pi}{2}} \left(\sqrt{v^2+4} - v \right) + \frac{v^2}{2} \\ &\leq \sqrt{\frac{\pi}{2}} |v| \left(\sqrt{1 + \frac{4}{M^2}} - 1 \right) + \frac{v^2}{2} \leq \frac{\sqrt{2\pi}|v|}{M} + \frac{v^2}{2} \leq 2v^2 + \frac{1}{M^2} \end{aligned}$$

using the elementary bound $|\sqrt{1+x^2} - 1| \leq |x|$ for all $x \geq 0$. When $v \geq -M$,

$$F(w,v) = \log \frac{\Psi(w)}{\Psi(v)} = \log\left(1 + \frac{\int_v^w e^{-\frac{t^2}{2}} \, \mathrm{d}t}{\int_{-\infty}^{v} e^{-\frac{t^2}{2}} \, \mathrm{d}t} \right) \leq \frac{\int_v^w e^{-\frac{t^2}{2}} \, \mathrm{d}t}{\int_{-\infty}^{v} e^{-\frac{t^2}{2}} \, \mathrm{d}t} \leq \frac{w - v}{\int_{-\infty}^{-M} e^{-\frac{t^2}{2}} \, \mathrm{d}t}$$

This completes the proof. □

Corollary 7.9. *Let $\Omega' \subset \mathbb{R}^d$ be open and bounded. Let μ' be a bounded, non-negative measure on Ω' and $\gamma > 0$. Define $\Psi(\cdot;\gamma)$ as in* (8)*. If $v \in L^2_{\mu'}$ then $\log \Psi(v;\gamma) \in L^1(\mu')$.*

ACCEPTED MANUSCRIPT

Proof. Lemma 7.8, and using that $\Psi(v;\gamma) = \Psi(v/\gamma;1)$, shows that $-\log\Psi(v,\gamma)$ grows quadratically as $v \to -\infty$. Note that $-\log\Psi(v,\gamma)$ asymptotes to zero as $v \to \infty$. Therefore $|\log\Psi(v,\gamma)| \le C(|v|^2+1)$ for some $C > 0$, which implies the claim. □

7.9. Weyl's Law

Lemma 7.10. *Let Ω and ρ satisfy Assumtptions 2–3 and let λ_k be the eigenvalues of $\mathcal{L}$ defined by* (4)*. Then, there exist positive constants c and C such that for all k large enough*

$$ck^{\frac{2}{d}} \le \lambda_k \le Ck^{\frac{2}{d}}.$$

Proof. Let B be a ball compactly contained in Ω and U a ball which compactly contains Ω. By assumptions on ρ for all $u \in H_0^1(B)\backslash\{0\}$

$$\frac{\int_B |\nabla u|^2 dx}{\int_B u^2 dx} \ge c_2 \frac{\int_\Omega |\nabla u|^2 \rho^2 dx}{\int_\Omega u^2 \rho dx}$$

where on RHS we consider the extension by zero of u to Ω. Therefore for any k-dimensional subspace V_k of $H_0^1(B)$

$$\max_{u\in V_k\backslash\{0\}} \frac{\int_B |\nabla u|^2 dx}{\int_B u^2 dx} \ge c_2 \max_{u\in V_k\backslash\{0\}} \frac{\int_\Omega |\nabla u|^2 \rho^2 dx}{\int_\Omega u^2 \rho dx}.$$

Consequently, using the Courant–Fisher characterization of eigenvalues,

$$\alpha_k = \inf_{\substack{V_k\subset H_0^1(B),\\ \dim V_k = k}} \max_{u\in V_k\backslash\{0\}} \frac{\int_B |\nabla u|^2 dx}{\int_B u^2 dx} \ge c_2 \inf_{\substack{V_k\subset H^1(\Omega),\\ \dim V_k = k}} \max_{u\in V_k\backslash\{0\}} \frac{\int_\Omega |\nabla u|^2 \rho^2 dx}{\int_\Omega u^2 \rho dx} = c_2\lambda_k$$

Since $\overline{\Omega}$ is an extension domain (as it has a Lipschitz boundary), there exists an bounded extension operator $E : H^1(\Omega) \to H_0^1(U)$. Therefore for some constant C_2 and all $u \in H^1(\Omega)$, $C_2 \int_\Omega |\nabla u|^2\rho^2 + u^2\rho dx \ge \int_U |\nabla Eu|^2 dx$. Arguing as above gives $C_2(\lambda_k + 1) \ge \beta_k$.

These inequalities imply the claim of the lemma, since the Dirichlet eigenvalues of the Laplacian on B, α_k satisfy $\alpha_k \le C_1 k^{\frac{2}{d}}$ for some C_1 and that Dirichlet eigenvalues of the Laplacian on U, β_k satisfy $\beta_k \ge c_1 k^{\frac{2}{d}}$ for some $c_1 > 0$. □

ACCEPTED MANUSCRIPT