Journal of Machine Learning Research 23 (2022) 1-28 Submitted 1/21; Revised 6/21; Published 3/22

Unlabeled Data Help in Graph-Based Semi-Supervised Learning: A Bayesian Nonparametrics Perspective

Daniel Sanz-Alonso SANZALONSO@UCHICAGO.EDU
Department of Statistics
University of Chicago
Chicago, IL 60637, USA

Ruiyi Yang YRY@UCHICAGO.EDU
Committee on Computational and Applied Mathematics
University of Chicago
Chicago, IL 60637, USA

Editor: Zhihua Zhang

Abstract

In this paper we analyze the graph-based approach to semi-supervised learning under a manifold assumption. We adopt a Bayesian perspective and demonstrate that, for a suitable choice of prior constructed with sufficiently many unlabeled data, the posterior contracts around the truth at a rate that is minimax optimal up to a logarithmic factor. Our theory covers both regression and classification.

Keywords: semi-supervised learning, graph-Laplacians, Bayesian nonparametrics, Gaussian fields on manifold

1. Introduction

Semi-supervised learning (SSL) refers to machine learning techniques that combine during training a small amount of labeled data with a large amount of unlabeled data. SSL has received increasing attention over the past decades because in many recent applications labeled data are expensive to collect while unlabeled data are abundant. Examples include analysis of body-worn videos and video surveillance (Qiao et al., 2019), text categorization and translation (Shi et al., 2010), image classification (Zhu, 2005) and protein structure prediction (Weston et al., 2005). The aim of this paper is to investigate whether unlabeled data can enhance learning performance. The answer to such question necessarily depends on the model assumptions and the methodology employed. Our main contribution is to show that under a standard manifold assumption, unlabeled data are helpful when using graph-based methods in a Bayesian setting. We do so by establishing that the optimal posterior contraction rate is achieved (up to a logarithmic factor) provided that the size of the unlabeled dataset grows sufficiently fast with the size of the labeled dataset.

The SSL problem of interest can be informally described as follows. Given labeled data $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ and unlabeled data $\{X_{n+1}, \dots, X_{N_n}\}$, the goal is to predict Y from X. More precisely, we are interested in the SSL problem of inferring the regression function $f_0(x) := \mathbb{E}(Y|X = x)$ at the given features $\mathcal{X}_{N_n} := \{X_1, \dots, X_{N_n}\}$ in either of these two settings:

1. Regression: $Y = f_0(X) + \eta$, where $\eta \sim \mathcal{N}(0, \sigma^2)$ with σ^2 known.

2. Binary classification: $\mathbb{P}(Y = 1|X) = f_0(X)$.

We analyze graph-based methods applied to this inference task under the *manifold assumption* that X takes values on a hidden manifold $\mathcal{M}$. This assumption is natural when the features $\mathcal{X}_{N_n}$ live in a high-dimensional ambient space but admit a low-dimensional representation (Bickel and Li, 2007; Niyogi, 2013; García Trillos et al., 2019b, 2020). For instance, Hein and Audibert (2005) and Costa and Hero (2006) demonstrate the low intrinsic dimension of standard datasets for image classification. Graph-based methods are well-suited under the manifold assumption as they allow to uncover the geometry of the hidden manifold and promote smoothness of the inferred f_0 along it. Indeed, graph-based methods are among the most powerful classification techniques for SSL problems where similar features are expected to belong to the same class (Belkin et al., 2004; Belkin and Niyogi, 2004; Zhu, 2005). The central idea behind traditional graph-based methods is to infer f_0 by minimizing an objective function comprising at least two terms: (i) a regularization term constructed with a graph-Laplacian of the features $\mathcal{X}_{N_n}$, which leverages the ability of unsupervised graph-based techniques to extract geometric information from $\mathcal{M}$; and (ii) a data-fidelity term that incorporates the labeled data. We adopt an analogous Bayesian perspective where: (i) the prior distribution $\Pi_n(\cdot \,|\, \mathcal{X}_{N_n})$ will be defined using a graph-Laplacian of the features $\mathcal{X}_{N_n}$ (hence the notation "given $\mathcal{X}_{N_n}$") to extract geometric information from $\mathcal{M}$; and (ii) the likelihood function promotes matching the labeled data. Assuming the data are independent so that the likelihood factorizes, the posterior takes the form

$$\Pi_n(B \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) \propto \int_B \prod_{i=1}^n L_{Y_i|X_i}(f)\, d\Pi_n(f \,|\, \mathcal{X}_{N_n}), \qquad B \in \mathcal{B}, \tag{1}$$

where $\mathcal{Y}_n := \{Y_1, \ldots, Y_n\}$ and $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}^{N_n}$ (here we are identifying functions over $\mathcal{X}_{N_n}$ with $\mathbb{R}^{N_n}$). The conditional likelihood of $Y_i|X_i$ is given by

$$L_{Y_i|X_i}(f) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|Y_i - f(X_i)|^2}{2\sigma^2}\right) \tag{2}$$

in the regression setting and

$$L_{Y_i|X_i}(f) = f(X_i)^{Y_i} \left[1 - f(X_i)\right]^{1-Y_i} \tag{3}$$

in the classification setting.

We shall analyze our Bayesian approach in a frequentist perspective by assuming the data is generated from a fixed truth f_0 and studying the contraction rate of the posterior around f_0, defined as the sequence of real numbers ε_n such that

$$\mathbb{E}_X \mathbb{E}_{f_0} \Pi_n \left(f : d_n(f, f_0) \geq M_n \varepsilon_n \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n\right) \xrightarrow{n\to\infty} 0,$$

for any sequence $M_n \to \infty$ and some suitable semi-metric d_n. Here the double expectation $\mathbb{E}_X \mathbb{E}_{f_0}$ is taken first with respect to the conditional data distribution of $\mathcal{Y}_n|X_1, \ldots, X_n$ specified by f_0 and then with respect to the marginal of $\mathcal{X}_{N_n}$. The idea of posterior contraction

rate was formally introduced in Ghosal et al. (2000) and Shen and Wasserman (2001) and has since then become a popular criterion for analyzing Bayesian methods. In particular, contraction of the posterior with rate ε_n gives a point estimator defined as

$$\widehat{f_n} = \arg\max_g \left[\Pi_n(f : d_n(f, g) \leq M_n \varepsilon_n \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n)\right]$$

that converges to f_0 with the same rate ε_n, which together with the minimax theory for statistical estimation quantifies the performance of our Bayesian SSL procedure.

Our main result, Theorem 1, shows that the posterior (1) contracts around f_0 at the minimax optimal rate (up to logarithmic factors), provided that the prior distribution $\Pi_n(\cdot \,|\, \mathcal{X}_{N_n})$ is carefully designed and that N_n grows at a certain polynomial rate with n. More broadly, our theory suggests that graph-based methods require abundant unlabeled data in order to effectively extract geometric information to regularize SSL problems.

1.1 Related Work

The question of whether unlabeled data enhance SSL performance has been widely studied. Positive and negative conclusions have been reached depending on the assumptions made on the relationship between the target function and the marginal data distribution, and also on the methodology employed (Cozman and Cohen, 2006; Bickel and Li, 2007; Liang et al., 2007; Wasserman and Lafferty, 2008; Niyogi, 2013; Singh et al., 2008). Two standard model assumptions are the *clustering assumption* (Seeger, 2000), which states that the target function is smooth over high density regions, and the *manifold assumption* that we have introduced. The latter has been extensively used in the statistical learning literature, and specifically in SSL, e.g. Bickel and Li (2007); Wasserman and Lafferty (2008); Castillo et al. (2014); Yang and Dunson (2016); García Trillos et al. (2019b, 2020).

Several methodologies for SSL have been developed based on generative modeling, support vector machines, semi-definite programming, graph-based methods, etc. (see the overview in Zhu (2005) and Chapelle et al. (2006)). Our focus is on graph-based approaches that combine label information with geometric information extracted from the unlabeled data employing graphical unsupervised techniques (Von Luxburg, 2007). This heuristic has motivated the use of graph-based regularizations in a wide number of applications, but a rigorous analysis of the mechanisms by which unlabeled data enhance the performance of graph-based SSL methods is only starting to emerge. The recent papers Bertozzi et al. (2021) and Hoffmann et al. (2020) studied posterior *consistency* (the asymptotics of the posterior probability $\Pi_n(f : d_n(f, f_0) \geq \varepsilon \,|\, \text{data})$ for a fixed small number ε) for a fixed sample size in the small noise limit, whereas we consider the large n limit and further establish posterior contraction rates. Rates of convergence for optimization rather than Bayesian formulations of graph-based SSL have been recently established in Calder et al. (2020). In a Bayesian framework, García Trillos and Sanz-Alonso (2018) and García Trillos et al. (2020) show the continuum limit of posterior distributions as the size of the unlabeled dataset grows, without increasing the size of the labeled dataset. These papers did not investigate whether the posteriors contract around the truth, and they did not demonstrate the value of unlabeled data in boosting learning performance. The work Kirichenko and van Zanten (2017) studies fully-supervised function estimation on large graphs assuming

that the truth changes with the size of the graph. In contrast, we investigate posterior contraction in a SSL setting with a *fixed* truth f_0 defined on the underlying manifold $\mathcal{M}$.

Our analysis uses tools from Bayesian nonparametrics and spectral analysis of graph-Laplacians. We will provide the necessary background on posterior contraction in Section 3, and we refer to (Ghosal and van der Vaart, 2017, Chapters 6, 8 and 11) for a comprehensive introduction to this subject. While numerous results on spectral convergence of graph-Laplacians can be found in the literature, our analysis of posterior contraction requires bounds in L^∞-metric with rates, recently developed in Dunson et al. (2021) and Calder et al. (2022). The idea of incorporating Bayesian nonparametrics tools in analyzing computational algorithms can also be found in Sanz-Alonso and Yang (2022b), which considers Gaussian process methodologies based on finite elements.

1.2 Main Contributions and Scope

Our main result, Theorem 1, is to our knowledge the first to establish posterior contraction rates for graph-based SSL. In doing so, we provide novel understanding on the relative value of labeled and unlabeled data, and we set forth a rigorous quantitative framework in which to analyze this question. We point out, however, two caveats. First, our theory is non-adaptive in the sense that the Bayesian methodology we analyze only achieves optimal contraction rates when a priori information on the smoothness of the truth is available. Our analysis is motivated by existing graph-based techniques, and the development and analysis of adaptive graph-based methods for SSL is an important research direction stemming from our work, but beyond the scope of this paper. Second, our results build on existing spectral convergence rates of graph-Laplacians that may be suboptimal; as a consequence, the required sample size of the unlabeled data that we establish may not be sharp. Due to the plug-in nature of our analysis, improved spectral convergence rates will automatically translate into a sharper bound on the sample complexity. Despite these caveats, our theory provides evidence that letting the unlabeled dataset grow polynomially with the labeled dataset, as suggested by Theorem 1, is a fundamental requirement for standard graph-based methods to achieve optimal contraction.

A central part of our proof is devoted to analyzing the convergence of a discretely indexed Gauss-Markov random field in an unstructured data cloud to a Matérn-type Gaussian field on $\mathcal{M}$. This is formalized in Theorem 7, a result that we believe to be of independent interest. Finally, our work contributes to the Bayesian nonparametrics literature on manifolds (Castillo et al., 2014).

Outline The rest of this paper is organized as follows. Section 2 introduces the construction of the graph-based prior and states our main result. Section 3 provides the necessary background on posterior contraction and outlines our analysis. Section 4 proves our main result, and we close in Section 5.

Notation We denote by $\mathcal{L}(Z)$ the law of the random variable Z. For a, b two real numbers, we denote $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. The symbol $\lesssim$ will denote less than or equal to up to a universal constant. For two real sequences $\{a_n\}$ and $\{b_n\}$, we denote (i) $a_n \ll b_n$ if $\lim_n(b_n/a_n) = 0$; (ii) $a_n = O(b_n)$ if $\limsup_n(b_n/a_n) \leq C$ for some positive

constant C; and (iii) $a_n \asymp b_n$ if $c_1 \leq \liminf_n(a_n/b_n) \leq \limsup_n(a_n/b_n) \leq c_2$ for some positive constants c_1, c_2. Finally we let γ denote the Lebesgue measure on $\mathbb{R}$.

2. Prior Construction and Main Result

In this section we introduce the graph-based prior $\Pi_n(\cdot \,|\, \mathcal{X}_{N_n})$ and we state the main result of this paper. Before doing so, we formalize our setting. We assume to be given labeled data $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} \mathcal{L}(X, Y)$ and unlabeled data $X_{n+1}, \ldots, X_{N_n} \overset{i.i.d.}{\sim} \mathcal{L}(X)$, where $\{X_i\}_{i=n+1}^{N_n}$ are independent from $\{X_i\}_{i=1}^{n}$. Recall that the goal is to estimate the regression function

$$f_0(x) = \mathbb{E}(Y|X = x)$$

at the given features $\mathcal{X}_{N_n} = \{X_1, \ldots, X_{N_n}\}$. We suppose that $\mu := \mathcal{L}(X)$ is supported on an m-dimensional smooth, connected, compact manifold $\mathcal{M}$ without boundary embedded in $\mathbb{R}^d$, with the absolute value of the sectional curvature bounded and with Riemannian metric inherited from the embedding. For technical reasons, we further assume that $m \geq 2$ and that $\mathcal{M}$ is a homogeneous manifold (the group of isometries acts transitively on $\mathcal{M}$, e.g. spheres and tori) normalized so that its volume is 1. We assume for simplicity that μ is the uniform distribution on $\mathcal{M}$. As discussed in Section 5, our results can be generalized to nonuniform marginal density.

2.1 Graph-based Prior

We now describe the construction of the graph-based prior $\Pi_n(\cdot \,|\, \mathcal{X}_{N_n})$ on f_0 restricted to the given features $\mathcal{X}_{N_n}$. The priors we consider have the general form

$$\Pi_n(\cdot \,|\, \mathcal{X}_{N_n}) = \mathcal{L}(\Phi(W_n) \,|\, \mathcal{X}_{N_n}), \tag{4}$$

where W_n is a Gauss-Markov random field in $\mathbb{R}^{N_n}$ whose covariance will be defined in terms of a graph-Laplacian (Von Luxburg, 2007) of $\mathcal{X}_{N_n}$. For the regression problem, Φ is taken to be the identity. For the classification problem, where f_0 takes values in $(0, 1)$, $\Phi : \mathbb{R} \to (0, 1)$ is a link function, which we assume throughout to be invertible and twice differentiable with $\Phi'/(\Phi(1-\Phi))$ uniformly bounded and $\int (\Phi'')^2/\Phi' \, d\gamma < \infty$. The logistic function, for instance, satisfies all these standard requirements.

In the remainder of this subsection we introduce and motivate our construction of the Gauss-Markov random field W_n. The starting point is to define a similarity matrix $H \in \mathbb{R}^{N_n \times N_n}$ whose entries $H_{ij} \geq 0$ represent the closeness between features X_i and X_j. For reasons discussed in Subsection 2.1.1 below, we set

$$H_{ij} := \frac{2(m+2)}{N_n \nu_m \zeta_{N_n}^{m+2}} \mathbf{1}\big\{|X_i - X_j| < \zeta_{N_n}\big\}, \tag{5}$$

where $|\cdot|$ is the Euclidean distance in $\mathbb{R}^d$, ν_m is the volume of the $m-$dimensional unit ball, and ζ_{N_n} is the *connectivity* of the graph, a user-specified parameter. Recall that a graph-Laplacian is a positive semi-definite matrix obtained by suitably transforming a similarity matrix. For concreteness, we work with the unnormalized graph-Laplacian matrix

$\Delta_{N_n} := D - H$, where D is the diagonal matrix with entries $D_{ii} = \sum_{j=1}^{N_n} H_{ij}$. For a vector $v \in \mathbb{R}^{N_n}$, naturally identified with a function on the data cloud $\mathcal{X}_{N_n}$, we have

$$v^T \Delta_{N_n} v = \frac{1}{2} \sum_{i=1}^{N_n} H_{ij} |v(X_i) - v(X_j)|^2. \tag{6}$$

We see that indeed Δ_{N_n} is positive semi-definite and that functions v that change slowly with respect to the similarity H have a small value of $v^T \Delta_{N_n} v$. This naturally suggests considering the Gaussian distribution $\mathcal{N}(0, (I_{N_n} + \Delta_{N_n})^{-s})$ since the likelihood for the samples will then have a term similar to (6). Based on this idea we shall define the Gauss-Markov random field W_n as the principal components of it. Precisely, let $\{(\lambda_i^{(N_n)}, \psi_i^{(N_n)})\}_{i=1}^{N_n}$ be the ordered eigenpairs of Δ_{N_n} and define

$$W_n = \sum_{i=1}^{k_{N_n}} \left[1 + \lambda_i^{(N_n)}\right]^{-\frac{s}{2}} \xi_i \psi_i^{(N_n)}, \qquad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{7}$$

where $k_{N_n} \ll N_n$ is to be determined. Notice that W_n as in (7) is a truncation of the Karhunen-Loève expansion of samples from $\mathcal{N}(0, (I_{N_n} + \Delta_{N_n})^{-s})$ and therefore the terminology "principal components".

The law of W_n depends on three parameters: the graph *connectivity* ζ_{N_n} used to define the similarity matrix H, the *smoothness* parameter s, and the principal components *truncation* parameter k_{N_n}. The connectivity ζ_{N_n} determines the sparsity of the precision matrix of W_n, and it should be taken to decrease with N_n to better resolve the geometry of $\mathcal{M}$ as more unlabeled data are available. The smoothness s controls the level of regularization. Larger s leads to faster decay of the coefficients in (7) and smoother samples. The truncation parameter $k_{N_n} \ll N_n$ allows us to keep only the components of the graph-Laplacian that contain useful geometric information on $\mathcal{M}$. Suitable choices and scalings of these three parameters, as well as further insights on their interpretation, will be given as we develop our theory. Importantly, the construction does not require knowledge of the underlying manifold $\mathcal{M}$, but only of its dimension. These data-driven Gauss-Markov fields have been used within various intrinsic approaches to Bayesian learning, see e.g. García Trillos and Sanz-Alonso (2018); García Trillos et al. (2019b, 2020); Harlim et al. (2020).

2.1.1 Interpretation as a Discretely Indexed Matérn Field

The Gauss-Markov field W_n can be interpreted as a discretely indexed Matérn Gaussian field (Sanz-Alonso and Yang, 2022a). Consider the Gaussian measure $\mathcal{N}(0, (I - \Delta)^{-s}))$, where $-\Delta$ denotes the Laplace-Beltrami operator on $\mathcal{M}$ and the fractional order operator is defined spectrally. Then, draws from $\mathcal{N}(0, (I - \Delta)^{-s}))$ admit a representation

$$W^{\mathcal{M}} = \sum_{i=1}^{\infty} (1 + \lambda_i)^{-\frac{s}{2}} \xi_i \psi_i, \qquad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{8}$$

where (λ_i, ψ_i)'s are the ordered eigenpairs of the Laplace-Beltrami operator. Note the analogy with (7). The field $W^{\mathcal{M}}$ is a generalization of the Matérn model to compact manifolds (Lindgren et al., 2011; Sanz-Alonso and Yang, 2022a). An important step in our

analysis of posterior contraction will be to show the convergence of W_n towards $W^{\mathcal{M}}$, in a sense to be made precise in Subsection 4.3, provided that the connectivity, smoothness and truncation parameters of W_n are suitably chosen. The similarity matrix defined in (5) enables this convergence result, but other choices (e.g. K-nearest neighbors) are possible. Key to showing convergence of W_n to $W^{\mathcal{M}}$ is the spectral convergence of Δ_{N_n} to $-\Delta$. The truncation parameter k_{N_n} is motivated by the fact that only the first eigenpairs of Δ_{N_n} accurately approximate those of $-\Delta$; see Proposition 10 below for a precise statement.

2.1.2 Reconciling Optimization and Bayesian Perspectives

To further motivate the definition of W_n, we relate our prior construction to the optimization literature. To streamline the discussion, let us focus on the regression problem where the link function Φ in (4) is the identity map. Classical graph-based optimization recovers f_0 at $\mathcal{X}_{N_n}$ using $f^T\Delta_{N_n}f$ as a regularizer and an appropriate data-fidelity term to match the labeled data, for instance,

$$\hat{f}_0 := \underset{f\in\mathbb{R}^{N_n}}{\arg\min} \sum_{i=1}^{n} |Y_i - f(X_i)|^2 + \lambda f^T\Delta_{N_n}f, \tag{9}$$

where λ controls the level of regularization. The solution to (9) is conceptually equivalent to the *maximum a posteriori* estimator in the Bayesian approach when the prior is chosen to be $\mathcal{N}(0, \Delta_{N_n}^{-1})$, where $\Delta_{N_n}^{-1}$ denotes the pseudo-inverse of Δ_{N_n}. The paper Nadler et al. (2009) shows that (9) is not well-posed when $m \geq 2$, in the sense that as the number of unlabeled data points increases the solution degenerates to a noninformative function. The authors suggest that this issue can be alleviated by defining the regularization term as $\lambda f^T\Delta_{N_n}^s f$ with $s > \frac{m}{2}$, so that higher order "derivatives" of f are controlled. Similar behavior has been observed with p-Laplacian regularizations (El Alaoui et al., 2016). The parameter s in (7) plays the exact same role, and we shall see that suitably scaling s with m is needed to warrant consistent learning in the limit of large unlabeled datasets.

2.2 Main Result

Now we are ready to state our main result. Let $\delta > 0$ be arbitrary. Let $p_m = \frac{3}{4}$ if $m = 2$ and $p_m = \frac{1}{m}$ otherwise. Let $\alpha_s = \frac{6m+6}{2s-3m+1}$ if $s \leq \frac{9}{2}m + \frac{5}{2}$ and $\alpha_s = 1$ otherwise.

Theorem 1 *Suppose $\Phi^{-1}(f_0) \in B^{\beta}_{\infty,\infty}$. Let Π_n be the prior defined by (4) and (7) with $s > \frac{3}{2}m - \frac{1}{2}$. Consider the following scaling for ζ_{N_n}, k_{N_n} and N_n.*

1. *$m \leq 4$: $\zeta_{N_n} \asymp N_n^{-\frac{1}{m+4+\delta}}(\log N_n)^{\frac{p_m}{2}}, k_{N_n} \asymp N_n^{\frac{m}{(m+4+\delta)(2s-3m+1)\alpha_s}}(\log N_n)^{-\frac{mp_m}{(4s-6m+2)\alpha_s}}$ with*

$$N_n \asymp n^{(m+4+\delta)\alpha_s}(\log n)^{\frac{(m+4+\delta)p_m}{2}}. \tag{10}$$

2. *$m \geq 5$: $\zeta_{N_n} \asymp N_n^{-\frac{1}{2m}}(\log N)^{\frac{p_m}{2}}, k_{N_n} \asymp N_n^{\frac{1}{(4s-6m+2)\alpha_s}}(\log N_n)^{-\frac{mp_m}{(4s-6m+2)\alpha_s}}$ with*

$$N_n \asymp n^{2m\alpha_s}(\log n)^{mp_m}. \tag{11}$$

Then, for ε_n a multiple of $n^{-\frac{(s-\frac{m}{2})\wedge\beta}{2s}}(\log n)^{\frac{(s-\frac{m}{2})\wedge\beta}{4s-2m}}$ and all $M_n \to \infty$,

$$\mathbb{E}_X \mathbb{E}_{f_0} \Pi_n \left(f \in L^\infty(\mu_{N_n}) : \|f - f_0\|_n \geq M_n \varepsilon_n \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n \right) \xrightarrow{n\to\infty} 0, \tag{12}$$

where μ_{N_n} is the empirical measure of $\mathcal{X}_{N_n}$ and $\|f - f_0\|_n^2 := \frac{1}{n}\sum_{i=1}^n |f(X_i) - f_0(X_i)|^2$. Here $\mathbb{E}_{f_0}$ denotes expectation with respect to the conditional distribution of $\mathcal{Y}_n | X_1, \ldots, X_n$ determined by f_0 and $\mathbb{E}_X$ denotes expectation with respect to the marginal distribution of $\mathcal{X}_{N_n}$.

Theorem 1 presents the posterior contraction rates with respect to the priors constructed in Section 2.1 under suitable choices of the parameters. Several remarks are in order.

The truth f_0 is assumed to belong to a Besov-type space $B^\beta_{\infty,\infty}$ on manifolds, defined in Subsection 4.2 in analogy to the usual Besov space $B^\beta_{\infty,\infty}(\mathbb{R}^m)$ on Euclidean space. We recall that $B^\beta_{\infty,\infty}(\mathbb{R}^m)$ coincides with the Hölder space $\mathcal{C}^\beta(\mathbb{R}^m)$ (the space of functions with $\lfloor\beta\rfloor$ continuous derivatives whose $\lfloor\beta\rfloor$-th derivative is $\beta - \lfloor\beta\rfloor$-Hölder continuous) when $\beta \notin \mathbb{N}$ and contains $\mathcal{C}^\beta(\mathbb{R}^m)$ when $\beta \in \mathbb{N}$ (see e.g. (Triebel, 1992, Theorem 1.3.4)). Therefore, $B^\beta_{\infty,\infty}$ can be interpreted as an analog of the Hölder space with parameter β on manifolds, and in particular represents a space of β-regular functions.

The key implication of Theorem 1 is that when $s - \frac{m}{2} = \beta$, we recover the rate $n^{-\beta/(2\beta+m)}(\log n)^{1/4}$, which is minimax optimal up to a logarithmic factor for β-regular functions. The assumption $s > \frac{3}{2}m - \frac{1}{2}$ then requires $\beta > m - \frac{1}{2}$, so optimal contraction rates can only be attained if f_0 is not too rough. Theorem 1 can be extended to hold for all $s > m$ if the eigenfunctions of the Laplace-Beltrami operator on $\mathcal{M}$ are uniformly bounded, which holds for the family of flat manifolds (Toth and Zelditch, 2002) that include for instance the tori. As mentioned in Subsection 1.2, the above choice of s requires knowing the regularity of f_0 and is not adaptive.

Another key ingredient of the result is the scaling for N_n as in (10) and (11), which are both larger than a multiple of n^{2m} since $\alpha_s \geq 1$ in all cases. In other words, the required sample size of the unlabeled data should grow polynomially with respect to the sample size of the labeled data in order to achieve the near optimal contraction rates described above, thereby justifying the claim that unlabeled data help.

We further remark that Theorem 1 only gives an upper bound for the required sample size N_n, whose proof (see Section 4) in fact has a plug-in nature. Suppose the sequence of Gaussian-Markov fields W_n in (7) converges in some semimetric d_n towards $W^{\mathcal{M}}$ in (8) with rate $\mathbb{E}_\xi d_n(W_n, W^{\mathcal{M}}) = R(N_n)$ for some function R. The required sample size is then obtained by matching $R(N_n)$ with n^{-1}. In particular, any improvement of the rate $R(N_n)$ shown in Subsection 4.3 will lead to a reduction of the required sample size. However, the spectral convergence rate in Proposition 10 and hence $R(N_n)$ should not be faster than the resolution of the point cloud, which is shown to scale like $N_n^{-1/m}$ in Proposition 6. Therefore by matching $N_n^{-1/m}$ with n^{-1} we expect a polynomial dependence of N_n on n to be necessary to achieve optimal contraction rate, further demonstrating the need of unlabeled data in our specific setting.

3. Posterior Contraction: Background and Set-up

In this section we present necessary background on posterior contraction rates. In Subsection 3.1 we review the main results on posterior contraction that our theory builds on, and in Subsection 3.2 we explain how these general results are used in our proof.

It will be important to note that our prior is constructed with the X_i's, and we shall view our observations as the Y_i's only, as the notation in (1) suggests. In particular, the Y_i's are independent but non-identically distributed (i.n.i.d.) and hence we will apply general results from Ghosal and van der Vaart (2007). This also explains the double expectation in (12), where the randomness of the X_i's is treated separately.

3.1 General Principles

Here we review general posterior contraction theory for the problem of estimating f_0 at the continuum level from i.n.i.d data, following Ghosal and van der Vaart (2007). Suppose the data $\{Y_i\}_{i=1}^n$ are generated according to a density $P_{f_0}^{(n)} = \prod_{i=1}^n p_{f_0,i}$ for some ground truth parameter f_0, where $p_{f_0,i}$ is the individual density for each observation. Let Π_n be a sequence of priors over f_0 that is supported on some parameter space $\mathcal{F}$ equipped with a semimetric d_n. Theorem 4 of Ghosal and van der Vaart (2007) states that

$$\mathbb{E}_{f_0}\Pi_n\left(f \in \mathcal{F} : d_n(f, f_0) \geq M_n\varepsilon_n \,|\, \mathcal{Y}_n\right) \xrightarrow{n\to\infty} 0, \tag{13}$$

provided that there exists sets $\mathcal{F}_n \subset \mathcal{F}$ and positive real numbers ε_n, K so that

$$\log N(\varepsilon_n, \mathcal{F}_n, d_n) \leq n\varepsilon_n^2, \tag{14}$$

$$\Pi_n(\mathcal{F}_n^c) \leq e^{-(K+4)n\varepsilon_n^2}, \tag{15}$$

$$\Pi_n(B_n^*(f_0, \varepsilon_n)) \geq e^{-Kn\varepsilon_n^2}, \tag{16}$$

where $N(\varepsilon_n, \mathcal{F}_n, d_n)$ is the minimum number of d_n-balls of radius ε_n needed to cover $\mathcal{F}_n$. Here

$$B_n^*(f_0, \varepsilon_n) := \left\{ f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n d_{\mathrm{KL}}(p_{f,i}, p_{f_0,i}) \leq \varepsilon_n^2, \quad \frac{1}{n}\sum_{i=1}^n v_{\mathrm{KL}}(p_{f,i}, p_{f_0,i}) \leq C\varepsilon_n^2 \right\}, \tag{17}$$

where C is a universal constant, $d_{\mathrm{KL}}(p, q) = \int p \log(p/q) d\gamma$ is the Kullback-Leibler divergence and $v_{\mathrm{KL}}(p, q) = \int p |\log(p/q) - d_{\mathrm{KL}}(p, q)|^2 d\gamma$. Conditions (15) and (14) roughly state that there are approximating sieves $\mathcal{F}_n$ which capture most of the prior probability while not being too large. Condition (16) requires sufficient prior mass near the truth f_0 and together with (14) further indicates that the priors should be "uniformly spread". In fact the three conditions are stronger than those in (Ghosal and van der Vaart, 2007, Theorem 4) but will suffice in our case for Gaussian process priors to be discussed shortly below. We shall refer to (Ghosal et al., 2000, Section 2) and (Ghosal and van der Vaart, 2007, Section 2) for further discussion on the interpretation and relaxation of these conditions.

The general theory can be used to analyze a wide range of statistical models but does not give a recipe for constructing the sieves $\mathcal{F}_n$ and ε_n. However, when the priors are Gaussian there exists a simple relation between ε_n and the priors. Suppose in addition that

$\Pi_n = \mathcal{L}(w_n)$ are Gaussian measures on some Banach space $(\mathbb{B}, \|\cdot\|_{\mathbb{B}})$ that converge to a fixed Gaussian measure $\Pi = \mathcal{L}(w)$ on the same Banach space with $10\mathbb{E}\|w_n - w\|_{\mathbb{B}}^2 \leq n^{-1}$. Theorem 2.2 in van der Vaart and van Zanten (2008a) then states that the contraction rate ε_n can be obtained by studying the *concentration function* of the limit prior Π, defined as

$$\varphi_{f_0}(\varepsilon) := \inf_{h \in \mathbb{H}: \|h - f_0\|_{\mathbb{B}} < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|w\|_{\mathbb{B}} < \varepsilon), \tag{18}$$

where $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is the *reproducing kernel Hilbert space* (RKHS) of Π (see e.g. van der Vaart and van Zanten (2008b) for more details on Gaussian measures and the associated RKHSs). More precisely, if ε_n satisfies $\varphi_{f_0}(\varepsilon_n) \leq n\varepsilon_n^2$, then for the same ε_n and some $\mathcal{F}_n$, the three conditions (14), (15) and (16) are satisfied (possibly for different proportion constants in front of $n\varepsilon_n^2$) with d_n and $B_n^*(f_0, \varepsilon_n)$ replaced by $\|\cdot\|_{\mathbb{B}}$ and $\{f : \|f - f_0\|_{\mathbb{B}} < \varepsilon_n\}$, respectively.

3.2 Application to Our Setting

In this subsection we discuss how we utilize the general theory outlined above, and provide a road map for the proof of Theorem 1. Notice that in (12) the sequence of posteriors are supported on the discrete space $L^\infty(\mu_{N_n})$, whose size changes with n. To alleviate this issue we will reduce the analysis to a sequence of posteriors $\Pi_n^{\mathcal{M}}(\cdot \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n)$ supported on the same continuum space $L^\infty(\mu)$ in Subsection 4.1. The corresponding sequence of priors will turn out to have the form $\Pi_n^{\mathcal{M}}(\cdot \,|\mathcal{X}_{N_n}) = \mathcal{L}(\Phi(\mathcal{I}W_n))$, with W_n defined in (7) and $\mathcal{I}$ a suitable interpolation map so that $\mathcal{I}W_n$ is a Gaussian process on $\mathcal{M}$ that approximates $W^{\mathcal{M}}$. Our analysis will then consist of the following steps. We will establish the three conditions (14), (15), (16) for $W^{\mathcal{M}}$ in Subsection 4.2, followed by a convergence rate analysis of $\mathcal{I}W_n$ towards $W^{\mathcal{M}}$ in Subsection 4.3, so that the three conditions are inherited by $\mathcal{I}W_n$. The assumptions on the link function Φ will then allow us to conclude similar conditions for $\Phi(\mathcal{I}W_n)$.

The discrepancy measure d_n can be taken as the empirical L^2-norm $\|\cdot\|_n$ in the regression case (Ghosal and van der Vaart, 2017, Section 8.3.2) and the root average square Hellinger distance in the classification setting (Ghosal and van der Vaart, 2007, Section 3), defined as

$$d_{n,H}^2(f, f') = \frac{1}{n} \sum_{i=1}^n \int \left(\sqrt{p_{f,i}} - \sqrt{p_{f',i}}\right)^2 d\gamma. \tag{19}$$

In the latter case, one can show that $d_{n,H}$ is equivalent to $\|\cdot\|_n$. Indeed, since the densities are uniformly bounded, $\|f - f'\|_n$ is bounded above by a multiple of (19). Furthermore (19) is upper bounded by a multiple of $\|f - f'\|_n$ by (Ghosal and van der Vaart, 2017, Lemma 2.8(ii)) given our assumption on Φ. Hence this explains the choice of $\|\cdot\|_n$ in (12).

A natural choice for the Banach space in our setting is then $\mathbb{B} = L^\infty(\mu)$ since (i) $d_n = \|\cdot\|_n \leq \|\cdot\|_{L^\infty(\mu)}$; and (ii) $B_n^*(f_0, \varepsilon_n) \supset \{f \in \mathcal{F} : \|f - f_0\|_{L^\infty(\mu)} \leq \tilde{C}\varepsilon_n\}$ for some universal constant $\tilde{C}$. The first point is clear and an upper bound on the metric entropy (14) in $\|\cdot\|_{L^\infty(\mu)}$ will automatically yield an upper bound in $\|\cdot\|_n$. The second point follows from the fact that one can upper bound the two quantities in (17) by $\|f - f_0\|_n$ and hence $\|f - f_0\|_{L^\infty(\mu)}$ in both the regression and classification setting. Therefore the prior mass condition (16) for $W^{\mathcal{M}}$ in $L^\infty(\mu)$ balls is sufficient to give the corresponding condition for

$B^*(f_0, \varepsilon_n)$. This motivates us to consider the continuum Gaussian field defined in (8) as an element in $L^\infty(\mu)$ in Subsection 4.2 and the $L^\infty(\mu)$ convergence rate in Subsection 4.3.

4. Proof of the Main Result

In this section we prove Theorem 1. An important part of the proof is to formalize the convergence of the Gauss-Markov random field W_n in (7), defined in the data cloud $\mathcal{X}_{N_n}$, to the Matérn field $W^{\mathcal{M}}$ in (8), defined in $\mathcal{M}$. To that end, we introduce an interpolation of W_n into $\mathcal{M}$

$$W_n^{\mathcal{M}} := \sum_{i=1}^{k_{N_n}} \left[1 + \lambda_i^{(N_n)}\right]^{-\frac{s}{2}} \xi_i \psi_i^{(N_n)} \circ T_{N_n}, \qquad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{20}$$

where $T_{N_n} : \mathcal{M} \to \{X_1, \ldots, X_{N_n}\}$ are transport maps to be specified in Proposition 6 below. In Subsection 4.1 we show that it suffices to establish posterior contraction with respect to the prior

$$\Pi_n^{\mathcal{M}}(\cdot \,|\, \mathcal{X}_{N_n}) := \mathcal{L}(\Phi(W_n^{\mathcal{M}}) \,|\, \mathcal{X}_{N_n}). \tag{21}$$

In Subsection 4.2 we show concentration properties of the limiting Gaussian field $W^{\mathcal{M}}$. In Subsection 4.3 we establish the $L^\infty(\mu)$-convergence rate of $W_n^{\mathcal{M}}$ towards $W^{\mathcal{M}}$. Combining the facts that $W_n^{\mathcal{M}}$ is close to $W^{\mathcal{M}}$ with the contraction properties of $W^{\mathcal{M}}$, we deduce that $\mathcal{L}(W_n^{\mathcal{M}} \,|\, \mathcal{X}_{N_n})$ satisfies the three conditions (14), (15), (16). We complete the proof in Subsection 4.4 by lifting these conditions to $\mathcal{L}(\Phi(W_n^{\mathcal{M}}) \,|\, \mathcal{X}_{N_n})$ in both the regression and classification problems.

4.1 Reduction via Interpolation

Here we show that in order to establish Theorem 1 it suffices to prove posterior contraction with respect to the continuum prior $\Pi_n^{\mathcal{M}}(\cdot \,|\, \mathcal{X}_{N_n})$ defined in (21). Let $\mathcal{I} := \mathcal{I}_{N_n} : L^\infty(\mu_{N_n}) \to L^\infty(\mu)$ be the interpolation map defined by $\mathcal{I}u := u \circ T_{N_n}$. Specifically, we claim that, in order to establish Theorem 1, it suffices to show that

$$\mathbb{E}_X \mathbb{E}_{f_0} \Pi_n^{\mathcal{M}} \left(f \in \mathcal{I}(L^\infty(\mu_{N_n})) : \|f - f_0\|_n \geq M_n \varepsilon_n \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n\right) \xrightarrow{n\to\infty} 0.$$

The only property of the maps T_{N_n} in (20) that we will use in this subsection is that $T_{N_n}(X_i) = X_i$. In order to establish the claim, let $A_n = \{f \in L^\infty(\mu_{N_n}) : \|f - f_0\|_n \geq M_n \varepsilon_n\}$. Since $A_n \subset \mathcal{I}^{-1}(\mathcal{I}(A_n))$ we have

$$\Pi_n(A_n \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) \leq \Pi_n(\mathcal{I}^{-1}(\mathcal{I}(A_n)) \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) = \mathcal{I}_\sharp[\Pi_n(\cdot \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n)](\mathcal{I}(A_n)),$$

where $\mathcal{I}_\sharp$ denotes push-forward through the map $\mathcal{I}$. (Recall that the push-forward of a measure ν through the map $\mathcal{I}$ is defined by the relationship $(\mathcal{I}_\sharp \nu)(B) := \nu(\mathcal{I}^{-1}(B))$.) Therefore, it follows from Lemma 2 below that

$$\Pi_n(A_n \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) \leq \Pi_n^{\mathcal{M}}(\mathcal{I}(A_n) | \mathcal{X}_{N_n}, \mathcal{Y}_n), \tag{22}$$

and hence it suffices to bound the right-hand side of (22). Since $T_{N_n}(X_i) = X_i$, we have

$$\begin{aligned}\mathcal{I}(A_n) = \{f \circ T_{N_n} : \|f - f_0\|_n \geq M_n\varepsilon_n\} &= \{f \circ T_{N_n} : \|f \circ T_{N_n} - f_0\|_n \geq M_n\varepsilon_n\} \\ &= \{f \in \mathcal{I}(L^\infty(\mu_{N_n})) : \|f - f_0\|_n \geq M_n\varepsilon_n\},\end{aligned}$$

and the claim is established.

Lemma 2 *It holds that* $\Pi_n^{\mathcal{M}}(\cdot \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) = \mathcal{I}_\sharp[\Pi_n(\cdot \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n)]$.

Proof Step 1: First we show that $\Pi_n^{\mathcal{M}}(\cdot\,|\mathcal{X}_{N_n}) = \mathcal{I}_\sharp[\Pi_n(\cdot\,|\mathcal{X}_{N_n})]$. Since $\mathcal{I}$ is linear, we have that $W_n^{\mathcal{M}} = \mathcal{I}W_n$. Then observe that $\mathcal{I}\big(\Phi(W_n)\big) = \Phi\big(\mathcal{I}(W_n)\big)$. Indeed, for $x \in T_{N_n}^{-1}(\{X_i\})$, we have

$$\mathcal{I}\big(\Phi(W_n)\big)(x) = \Phi(W_n)(X_i) = \Phi\big(W_n(X_i)\big) = \Phi\big(\mathcal{I}(W_n)(x)\big) = \Phi\big(\mathcal{I}(W_n)\big)(x).$$

Therefore, we have that

$$\Pi_n^{\mathcal{M}}(\cdot\,|\mathcal{X}_{N_n}) = \mathcal{L}\big(\Phi(W_n^{\mathcal{M}})|\mathcal{X}_{N_n}\big) = \mathcal{L}\big(\mathcal{I}\big(\Phi(W_n)\big)|\mathcal{X}_{N_n}\big).$$

Finally, note that $\mathcal{I}_\sharp[\Pi_n(\cdot\,|\mathcal{X}_{N_n})] = \mathcal{L}\big(\mathcal{I}\big(\Phi(W_n)\big)|\mathcal{X}_{N_n}\big)$, since

$$\mathcal{I}_\sharp\Pi_n(B|\mathcal{X}_{N_n}) = \Pi_n\big(\mathcal{I}^{-1}(B)|\mathcal{X}_{N_n}\big) = \mathbb{P}\big(\Phi(W_n) \in \mathcal{I}^{-1}(B)|\mathcal{X}_{N_n}\big) = \mathbb{P}\big(\mathcal{I}(\Phi(W_n)) \in B|\mathcal{X}_{N_n}\big).$$

Step 2: Now we show that $\Pi_n^{\mathcal{M}}(\cdot \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) = \mathcal{I}_\sharp[\Pi_n(\cdot \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n)]$. By definition of pushforward measure, it suffices to show that, for any measurable B,

$$\Pi_n^{\mathcal{M}}(B \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) = \Pi_n(\mathcal{I}^{-1}(B) \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n).$$

Using Step 1, the left-hand side equals

$$\Pi_n^{\mathcal{M}}(B \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n) = \frac{\int_B \prod_{i=1}^n L_{Y_i|X_i}(f)\, d\mathcal{I}_\sharp[\Pi_n(f \,|\, \mathcal{X}_{N_n})]}{\int_{\mathcal{I}(L^\infty(\mu_{N_n}))} \prod_{i=1}^n L_{Y_i|X_i}(f)\, d\mathcal{I}_\sharp[\Pi_n(f \,|\, \mathcal{X}_{N_n})]}, \tag{23}$$

where $L_{Y_i|X_i}(f)$ is given in (2) and (3). Note that pointwise values of f are well-defined since $\mathcal{I}_\sharp\Pi_n$ is supported on $\mathcal{I}(L^\infty(\mu_{N_n}))$. By the change-of-variable formula for pushforward measures,

$$(23) = \frac{\int_{\mathcal{I}^{-1}(B)} \prod_{i=1}^n L_{Y_i|X_i} \circ \mathcal{I}(f_n)\, d\Pi_n(f_n \,|\, \mathcal{X}_{N_n})}{\int_{L^\infty(\mu_{N_n})} \prod_{i=1}^n L_{Y_i|X_i} \circ \mathcal{I}(f_n)\, d\Pi_n(f_n \,|\, \mathcal{X}_{N_n})},$$

which equals (1) with B replaced by $\mathcal{I}^{-1}(B)$, by noticing that $L_{Y_i|X_i} \circ \mathcal{I}(f_n)$ is exactly the same as in (2) and (3). The result follows. ■

4.2 Regularity and Contraction Properties of the Limiting Field

In this subection we study the limit Gaussian measure $\pi = \mathcal{N}(0, (I-\Delta)^{-s}) = \mathcal{L}(W^{\mathcal{M}})$. Recall that the samples admit Karhunen-Loève expansion

$$W^{\mathcal{M}} = \sum_{i=1}^{\infty} (1+\lambda_i)^{-\frac{s}{2}} \xi_i \psi_i, \qquad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \tag{24}$$

where (λ_i, ψ_i)'s are the eigenpairs of the Laplace-Beltrami operator $-\Delta$ on $\mathcal{M}$. Notice that a larger s leads to a faster decay of the coefficients and hence more regular sample paths. From Weyl's law that $\lambda_i \asymp i^{\frac{2}{m}}$ (see e.g. (Canzani, 2013, Theorem 72)), setting $s > \frac{m}{2}$ makes π a well-defined measure on $L^2(\mu)$. To ensure almost sure continuity of the samples, so that point evaluations are well-defined and π is supported over $L^\infty(\mu)$, we need the stronger assumption that $s > m - \frac{1}{2}$.

Lemma 3 *If $s > m - \frac{1}{2}$, then samples of π are almost surely continuous.*

Proof By (Lang et al., 2016, Corollary 4.5), it suffices to show that

$$\mathbb{E}_\xi |W^{\mathcal{M}}(x) - W^{\mathcal{M}}(y)|^2 \le C d_{\mathcal{M}}(x,y)^\eta,$$

for some $\eta > 0$. We have

$$\begin{aligned}
\mathbb{E}_\xi |W^{\mathcal{M}}(x) - W^{\mathcal{M}}(y)|^2 &= \mathbb{E}_\xi \left(\sum_{i=1}^{\infty} (1+\lambda_i)^{-\frac{s}{2}} \xi_i (\psi_i(x) - \psi_i(y)) \right)^2 \\
&= \sum_{i=1}^{\infty} (1+\lambda_i)^{-s} |\psi_i(x) - \psi_i(y)|^2,
\end{aligned}$$

since $\mathbb{E}_\xi \xi_i \xi_j = 0$ for $i \neq j$. By Proposition 13,

$$\begin{aligned}
&|\psi_i(x) - \psi_i(y)| \le C \lambda_i^{\frac{m-1}{4}}, \\
&|\psi_i(x) - \psi_i(y)| \le \|\nabla \psi_i\|_{L^\infty(\mu)} d_{\mathcal{M}}(x,y) \le C \lambda_i^{\frac{m+1}{2}} d_{\mathcal{M}}(x,y).
\end{aligned}$$

Using Weyl's law, we have

$$\begin{aligned}
\mathbb{E}_\xi |W^{\mathcal{M}}(x) - W^{\mathcal{M}}(y)|^2 &\lesssim \sum_{i=1}^{\infty} (1+\lambda_i)^{-s} \min \left\{ \lambda_i^{m+1} d_{\mathcal{M}}(x,y)^2, \lambda_i^{\frac{m-1}{2}} \right\} \\
&\lesssim \sum_{i=1}^{\infty} i^{-\frac{2s}{m}} \min \left\{ i^{\frac{2m+2}{m}} d_{\mathcal{M}}(x,y)^2, i^{\frac{m-1}{m}} \right\} \\
&\lesssim d_{\mathcal{M}}(x,y)^2 \int_1^K z^{\frac{-2s+2m+2}{m}} dz + \int_K^\infty z^{\frac{-2s+m-1}{m}} dz.
\end{aligned}$$

Setting $K = d_{\mathcal{M}}(x,y)^{-\frac{2m}{m+3}}$, we have

$$\mathbb{E}_\xi |W^{\mathcal{M}}(x) - W^{\mathcal{M}}(y)|^2 \lesssim d_{\mathcal{M}}(x,y)^{\frac{4s-4m+2}{m+3}}.$$

The result follows since $s > m - \frac{1}{2}$. ■

Remark 4 *The argument above relies on the worst case $L^\infty(\mu)$ bound on the eigenfunctions of $-\Delta$ given in Proposition 13. If the eigenfunctions are uniformly bounded (which holds for the family of flat manifolds (Toth and Zelditch, 2002) that includes, for instance, the torus), then continuity can be guaranteed for $s > \frac{m}{2}$.*

From now on, we shall consider π as a measure over $L^\infty(\mu)$ with continuous sample paths, where pointwise evaluation is well-defined. Using the series representation (24), the reproducing kernel Hilbert space $\mathbb{H}$ associated with π has the following characterization

$$\mathbb{H} = \left\{ h = \sum_{i=1}^{\infty} h_i \psi_i : \|h\|_{\mathbb{H}}^2 := \sum_{i=1}^{\infty} h_i^2 (1+\lambda_i)^s < \infty \right\}. \tag{25}$$

From the general theory in van der Vaart and van Zanten (2008a), the concentration properties of W can be characterized by the concentration function

$$\varphi_{w_0}(\varepsilon) := \inf_{h \in \mathbb{H}: \|h - w_0\|_{L^\infty(\mu)} < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \mathbb{P}(\|W\|_{L^\infty(\mu)} < \varepsilon), \tag{26}$$

where w_0 belongs to the support of π. For our purpose, w_0 will be set as $\Phi^{-1}(f_0)$. The three conditions (14), (15) and (16) hold for the ε_n that satisfies $\varphi_{w_0}(\varepsilon_n) \le n\varepsilon_n^2$.

Before stating our main result in this subsection, we follow Castillo et al. (2014) and Coulhon et al. (2012) to define a Besov space which will be needed to characterize the regularity of w_0. Let Ψ be an even function in the Schwartz space $\mathcal{S}(\mathbb{R})$ with

$$0 \le \Psi \le 1, \quad \Psi \equiv 1 \text{ on } [-\frac{1}{2}, \frac{1}{2}], \quad \text{supp}(\Psi) \subset [-1, 1].$$

Define the Besov space

$$B_{\infty,\infty}^{\beta} := \left\{ w : \|w\|_{B_{\infty,\infty}^{\beta}}^{\beta} := \sup_{j \in \mathbb{N}} 2^{\beta j} \|\Psi_j(\sqrt{\Delta}) w(\cdot) - w(\cdot)\|_{L^\infty(\mu)} < \infty \right\},$$

where $\Psi_j(\cdot) = \Psi(2^{-j}\cdot)$ and, for $w = \sum_{i=1}^{\infty} w_i \psi_i$,

$$\Psi_j(\sqrt{\Delta}) w := \sum_{i=1}^{\infty} \Psi_j(\sqrt{\lambda_i}) w_i \psi_i.$$

It was shown in (Coulhon et al., 2012, Proposition 6.2) that the definition is independent of the choice of Ψ.

Theorem 5 *Suppose $w_0 \in B_{\infty,\infty}^{\beta}$. Consider the prior $\pi = \mathcal{N}(0, (I-\Delta)^{-s})$ with $s > \beta \wedge m - \frac{1}{2}$. Then there exists sets $B_n \subset L^\infty(\mu)$ so that*

$$\begin{aligned} \log N(3\varepsilon_n, B_n, \|\cdot\|_{L^\infty(\mu)}) &\le 6Cn\varepsilon_n^2, \\ \pi(B_n^c) &\le e^{-Cn\varepsilon_n^2}, \\ \pi(\|w - w_0\|_{L^\infty(\mu)} < 2\varepsilon_n) &\ge e^{-n\varepsilon_n^2}, \end{aligned}$$

where ε_n is a multiple of $n^{-\frac{(s-\frac{m}{2})\wedge\beta}{2s}} (\log n)^{\frac{(s-\frac{m}{2})\wedge\beta}{4s-2m}}$ and $C > 1$ is a constant satisfying $e^{-Cn\varepsilon_n^2} < \frac{1}{2}$.

Proof As noted above, it suffices to show that the above ε_n satisfies $\varphi_{w_0}(\varepsilon_n) \le n\varepsilon_n^2$, where φ_{w_0} is defined in (26). To bound the small ball probability, it suffices by general results from Li and Linde (1999) to bound the metric entropy of $\mathbb{H}_1$, the unit ball in $\mathbb{H}$. Notice that (25) has the form of a Sobolev ball of regularity s. Classical results for Sobolev spaces on $\mathbb{R}^m$ such as (Edmunds and Triebel, 1996, Theorem 3.3.2) have shown that the entropy in the L^∞ metric is bounded by a multiple of $\varepsilon^{-m/s}$. The manifold case has been treated in (Kushpel and Levesley, 2015, Theorem 3.12) assuming homogeneity, which includes an additional logarithmic factor, i.e.,

$$\log N(\varepsilon, \mathbb{H}_1, \|\cdot\|_{L^\infty(\mu)}) \lesssim \varepsilon^{-\frac{m}{s}} \left(\log \frac{1}{\varepsilon}\right)^{\frac{1}{2}}. \tag{27}$$

The original theorem was about entropy numbers but can be translated to the above statement on metric entropy and its proof suggested that when $p = 2$ the theorem holds with $s > \frac{m}{2}$. In particular their Sobolev class is defined as

$$I_s U_2 := \left\{ h = h_1 + \sum_{i=2}^{\infty} \lambda_i^{-\frac{s}{2}} h_i \psi_i : \sum_{i=1}^{\infty} h_i^2 \le 1 \right\},$$

which is compatible with our $\mathbb{H}$. (We remark that although Kushpel and Levesley (2015) defined their Sobolev class with a further mean-zero condition, their proof actually applied to the general case.) Now by (27) and (Li and Linde, 1999, Theorem 1.2) we have

$$-\log \mathbb{P}(\|w\|_{L^\infty(\mu)} < \varepsilon) \lesssim \varepsilon^{-\frac{2m}{2s-m}} \left(\log \frac{1}{\varepsilon}\right)^{\frac{s}{2s-m}}. \tag{28}$$

For the decentering function, let $C_0 = \|w_0\|_{B^\beta_{\infty,\infty}}$ and consider $h = \Psi_J(\sqrt{\Delta})w_0$ with J large enough so that $C_0 2^{-\beta J} \le \varepsilon$. Since $w_0 \in B^\beta_{\infty,\infty}$, we have

$$\|\Psi_j(\sqrt{\Delta})w_0 - w_0\|_{L^\infty(\mu)} \le C_0 2^{-\beta j}, \qquad j \in \mathbb{N}. \tag{29}$$

In particular $\|h - w_0\|_{L^\infty(\mu)} \le C_0 2^{-\beta J} \le \varepsilon$. Suppose w_0 has the representation $w_0 = \sum_{i=1}^\infty w_i \psi_i$. We then have $h = \sum \Psi_J(\sqrt{\lambda_i}) w_i \psi_i \in \mathbb{H}$ since $\Psi_J(\sqrt{\lambda_i}) = 0$ for $\sqrt{\lambda_i} > 2^J$. Moreover, since $\Psi_J \le 1$,

$$\|h\|_{\mathbb{H}}^2 \le \sum_{\sqrt{\lambda_i} \le 2^J} w_i^2 (1+\lambda_i)^s \le \sum_{\sqrt{\lambda_i} \le 1} 2^s w_i^2 + \sum_{j=1}^{J} \sum_{2^{j-1} < \sqrt{\lambda_i} \le 2^j} w_i^2 (1+\lambda_i)^s. \tag{30}$$

By (29), we have

$$\begin{aligned} \sum_{2^{j-1} < \sqrt{\lambda_i} \le 2^j} w_i^2 (1+\lambda_i)^s \lesssim 2^{2js} \sum_{2^{j-1} < \sqrt{\lambda_i}} w_i^2 &\le 2^{2js} \|\Psi_{j-1}(\sqrt{\Delta})w_0 - w_0\|_2^2 \\ &\lesssim 2^{2js} \|\Psi_{j-1}(\sqrt{\Delta})w_0 - w_0\|_\infty^2 \lesssim 2^{2(s-\beta)j}. \end{aligned}$$

Therefore recalling that $C_0 2^{-\beta J} \leq \varepsilon$,

$$(30) \lesssim \sum_{\sqrt{\lambda_i} \leq 1} 2^s w_i^2 + \sum_{j=1}^{J} 2^{2(s-\beta)j} \lesssim 2^{2(s-\beta)J} \lesssim \varepsilon^{-\frac{2(s-\beta)}{\beta}},$$

since the first sum remains bounded as $J \to \infty$. Combining the above with (28), we get

$$\varphi_{w_0}(\varepsilon) \lesssim \varepsilon^{-\frac{2m}{2s-m}} \left(\log \frac{1}{\varepsilon}\right)^{\frac{s}{2s-m}} + \varepsilon^{-\frac{2(s-\beta)}{\beta}},$$

which implies

$$\frac{\varphi_{w_0}(\varepsilon)}{\varepsilon^2} \lesssim \varepsilon^{-\frac{2s}{s-\frac{m}{2}}} \left(\log \frac{1}{\varepsilon}\right)^{\frac{s}{2s-m}} + \varepsilon^{-\frac{2s}{\beta}}.$$

Therefore by choosing

$$\varepsilon_n = C n^{-\frac{(s-\frac{m}{2}\wedge\beta)}{2s}} (\log n)^{\frac{(s-\frac{m}{2})\wedge\beta}{4s-2m}}$$

for a large enough constant C, the condition $\varphi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2$ is satisfied. The three assertions then follow from (van der Vaart and van Zanten, 2008a, Theorem 2.1). ■

4.3 Convergence of Gaussian Fields in $L^\infty(\mu)$

In this subsection we establish $L^\infty(\mu)$ convergence of $W_n^{\mathcal{M}}$ to $W^{\mathcal{M}}$. We shall denote $N := N_n$ throughout the rest of this subsection to simplify our notation. Recall that

$$\begin{aligned} W_n^{\mathcal{M}} &= \sum_{i=1}^{k_N} \left[1 + \lambda_i^{(N)}\right]^{-\frac{s}{2}} \xi_i \psi_i^{(N)} \circ T_N, \qquad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \\ W^{\mathcal{M}} &= \sum_{i=1}^{\infty} (1+\lambda_i)^{-\frac{s}{2}} \xi_i \psi_i, \qquad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1), \end{aligned} \tag{31}$$

In order to show the convergence, the transportation maps T_N will be assumed to be close to the identity in the sense made precise in the following proposition from (García Trillos et al., 2019a, Theorem 2).

Proposition 6 *For $\gamma > 1$, there exists a transportation map $T_N : \mathcal{M} \to \{X_1, \ldots, X_N\}$ satisfying $T_{N_n}(X_i) = X_i$ so that, with probability $1 - O(N^{-\gamma})$,*

$$\rho_N := \sup_{x \in \mathcal{M}} d_{\mathcal{M}}(x, T_N(x)) \lesssim \frac{(\log N)^{p_m}}{N^{1/m}}, \tag{32}$$

where $p_m = \frac{3}{4}$ if $m = 2$ and $p_m = \frac{1}{m}$ otherwise.

The transport map T_{N_n} is a measure-preserving transformation in the sense that $\mu(T_{N_n}^{-1}(U)) = \mu_{N_n}(U)$ for all $U \subset \mathcal{M}$ measurable. The additional requirement $T_N(X_i) = X_i$ that was absent in (García Trillos et al., 2019a, Theorem 2) is valid here since modifying T_N on a set of μ-measure 0 does not affect the measure preserving property and (32) still holds in this case. The scaling in (32) can be thought as the resolution of the point cloud, and is important in suitably defining the choice of connectivity ζ_N, as we shall see.

The main result of this subsection is the following.

Theorem 7 *Let $\delta > 0$ be arbitrary.*

1. *$\frac{3}{2}m - \frac{1}{2} < s \leq \frac{9}{2}m + \frac{5}{2}$. Set*

$$\begin{cases} \zeta_N \asymp N^{-\frac{1}{m+4+\delta}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{1}{(m+4+\delta)(6+\frac{6}{m})}} (\log N)^{-\frac{mp_m}{12m+12}}, & m \leq 4, \\ \zeta_N \asymp N^{-\frac{1}{2m}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{1}{12m+12}} (\log N)^{-\frac{mp_m}{12m+12}}, & m \geq 5. \end{cases}$$

Then, with probability tending to 1,

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim \begin{cases} N^{-\frac{2s-3m+1}{(m+4+\delta)(6m+6)}} (\log N)^{\frac{p_m(2s-3m+1)}{12m+12}}, & m \leq 4, \\ N^{-\frac{2s-3m+1}{m(12m+12)}} (\log N)^{\frac{p_m(2s-3m+1)}{12m+12}}, & m \geq 5. \end{cases}$$

2. *$s > \frac{9}{2}m + \frac{5}{2}$. Set*

$$\begin{cases} \zeta_N \asymp N^{-\frac{1}{m+4+\delta}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{m}{(m+4+\delta)(2s-3m+1)}} (\log N)^{-\frac{mp_m}{4s-6m+2}}, & m \leq 4, \\ \zeta_N \asymp N^{-\frac{1}{2m}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{1}{4s-6m+2}} (\log N)^{-\frac{mp_m}{4s-6m+2}}, & m \geq 5. \end{cases}$$

Then, with probability tending to 1,

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim \begin{cases} N^{-\frac{1}{m+4+\delta}} (\log N)^{\frac{p_m}{2}}, & m \leq 4, \\ N^{-\frac{1}{2m}} (\log N)^{\frac{p_m}{2}}, & m \geq 5. \end{cases}$$

We remark that the statement "with probability tending to 1" refers to the randomness coming from the X_i's. By solving for N so that the rate matches n^{-1}, we get the following.

Corollary 8 *Let $\delta > 0$ be arbitrary. Let*

$$N \asymp n^{\alpha_1} (\log n)^{\alpha_2},$$

where

$$\begin{cases} \alpha_1 = \frac{(m+4+\delta)(6m+6)}{2s-3m+1}, \quad \alpha_2 = \frac{p_m(m+4+\delta)}{2}, & \text{if} \quad m \leq 4 \ , \ \frac{3}{2}m - \frac{1}{2} < s \leq \frac{9}{2}m + \frac{5}{2}, \\ \alpha_1 = \frac{m(12m+12)}{2s-3m+1}, \quad \alpha_2 = mp_m, & \text{if} \quad m \geq 5 \ , \ \frac{3}{2}m - \frac{1}{2} < s \leq \frac{9}{2}m + \frac{5}{2}, \\ \alpha_1 = m + 4 + \delta, \quad \alpha_2 = \frac{p_m(m+4+\delta)}{2}, & \text{if} \quad m \leq 4 \ , \ s > \frac{9}{2}m + \frac{5}{2}, \\ \alpha_1 = 2m, \quad \alpha_2 = mp_m, & \text{if} \quad m \geq 5 \ , \ s > \frac{9}{2}m + \frac{5}{2}. \end{cases}$$

Let ζ_N and k_N have the same scaling as in Theorem 7. Then, with probability tending to 1,

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim n^{-1}.$$

Remark 9 *For flat manifolds the results of Theorem 7 and Corollary 8 can be shown to hold for $s > m$ with corresponding modifications in the scaling of the parameters.*

The key to show the above results is to derive convergence rates for

$$|\lambda_i^{(N)} - \lambda_i|, \qquad \text{and} \qquad \|\psi_i^{(N)} \circ T_N - \psi_i\|_{L^\infty(\mu)}.$$

We shall build our analysis on the existing results from the literature. Recall that ζ_N is the connectivity of the graph and the resolution ρ_N defined in (32). Assuming we are in the event that (32) holds, the following results from (Sanz-Alonso and Yang, 2022a, Theorem 4.6 & 4.7) bound the spectral approximations.

Proposition 10 (Spectral Approximation) *Suppose $\rho_N \ll \zeta_N$ and $\zeta_N \sqrt{\lambda_{k_N}} \ll 1$ for N large. Then there exists orthonormalized eigenfunctions $\{\psi_i^{(N)}\}_{i=1}^N$ and $\{\psi_i\}_{i=1}^\infty$ so that, for $i = 1, \dots, k_N$,*

$$|\lambda_i^{(N)} - \lambda_i| \lesssim \lambda_i \left(\frac{\rho_N}{\zeta_N} + \zeta_N \sqrt{\lambda_i} \right), \tag{33}$$

$$\|\psi_i^{(N)} \circ T_N - \psi_i\|_{L^2(\mu)}^2 \lesssim i^3 \left(\frac{\rho_N}{\zeta_N} + \zeta_N \sqrt{\lambda_i} \right). \tag{34}$$

Remark 11 *Proposition 10 bounds the eigenfunction approximation in $L^2(\mu)$ norm and we need to lift such result to the $L^\infty(\mu_N)$ norm. We remark that $L^\infty(\mu_N)$ convergence rates have already been established in the literature (see e.g. Dunson et al. (2021) and Calder et al. (2022)). However these results do not contain an explicit proportion constant in terms of the index i as in* (34)*. Instead of going through their details to find the explicit constants, which would make our presentation much more involved, we directly show $L^\infty(\mu_N)$ bounds based on Proposition 10, by following the same idea as in Calder et al. (2022). Since we build our results from* (34)*, which is not the sharpest bound in the literature, our results in Theorem 7 and Corollary 8 suffer the same drawback. Nevertheless, the goal of this paper is to demonstrate the idea that unlabeled data helps and hence the finding sharpest scaling of N_n is less essential.*

Below we record four results from (Calder and García Trillos, 2022, Theorem 3.3), (Donnelly, 2006, Theorem 1.2) and (Xu, 2006, Equation (2.10)), (Calder et al., 2020, Corollary 2.5) that will be needed.

Proposition 12 (Pointwise Error of Δ_N) *Let $f \in \mathcal{C}^3(\mathcal{M})$. Then with probability $1 - 2n \exp(-cN\zeta_N^{m+4})$,*

$$\|\Delta_N f(x) - \Delta f(x)\|_{L^\infty(\mu_N)} \leq C(1 + \|f\|_{\mathcal{C}^3(\mathcal{M})})\zeta_N.$$

Proposition 13 (Bounds on Eigenfunctions and Their Gradients) *Let ψ be an $L^2(\mu)$-normalized eigenfunction of $-\Delta$ associated with eigenvalue $\lambda \neq 0$. Then, for $k \in \mathbb{N}$,*

$$\|\psi\|_{L^\infty(\mu)} \leq C\lambda^{\frac{m-1}{4}},$$
$$\|\nabla^k \psi\|_{L^\infty(\mu)} \leq C\lambda^{k+\frac{m-1}{2}}.$$

Proposition 14 (L^∞ Bound In Terms of L^1 Bound) *Suppose $\zeta_N \le \frac{c}{\Lambda+1}$ where c is a sufficiently small constant depending only on $\mathcal{M}$. Then with probability $1-C\zeta_N^{-6m}\exp(-cN\zeta_N^{m+4})-2N\exp(-cN(\Lambda+1)^{-m})$*

$$\|u\|_{L^\infty(\mu_N)} \le C(\Lambda+1)^{m+1}\|u\|_{L^1(\mu_N)}$$

for all $u \in L^2(\mu_N)$ with $\lambda_u < \Lambda$, where

$$\lambda_u := \frac{\|\Delta_N u\|_{L^\infty(\mu_N)}}{\|u\|_{L^\infty(\mu_N)}}. \tag{35}$$

Notice that the high probability condition is satisfied only if

$$\zeta_N^{-6m}\exp(-cN\zeta_N^{m+4}) - 2N\exp(-cN(\Lambda+1)^{-m}) \to 0$$

and this requires some care in setting the scaling of ζ_N and k_N. In particular a sufficient condition for ζ_N is $\zeta_N \gtrsim N^{-\frac{1}{m+4+\delta}}$ for an arbitrarily small $\delta > 0$. On the other hand the condition for Λ reduces to the scaling of k_N. Indeed if $u = \psi_i^{(N)}$, then $\lambda_u = \lambda_i^{(N)}$ and therefore Λ can be chosen as a multiple of λ_i given the spectral approximation in Proposition 10. We will show that the same choice of Λ suffices when $u = \psi_i^{(N)} - \psi_i$. Since we are interested in bounding the first k_N eigenfunctions, we can set Λ to be multiple of λ_{K_N}. Therefore in order to make $2N\exp\left(-cN(\Lambda+1)^{-m}\right) \ll 1$, it is sufficient to have $\lambda_{k_N} \lesssim N^{\frac{1-\delta}{m}}$, i.e., if $k_N \lesssim N^{\frac{1-\delta}{2}}$ by Weyl's law. Therefore we should keep the following in mind when we set the scaling later

$$N^{-\frac{1}{m+4+\delta}} \lesssim \zeta_N, \qquad k_N \lesssim N^{\frac{1-\delta}{m}}, \qquad \zeta_N k_N^{\frac{2}{m}} \lesssim 1, \tag{36}$$

where the third requirement corresponds to the condition that $\zeta_N(\Lambda+1) \le c$.

Now we are ready to show the $L^\infty(\mu_N)$ bound of eigenfunction approximations.

Lemma 15 *Under the same conditions and the intersection of the high probability events as in Propositions 10 and 14 we have, for $i = 1, \dots, k_N$,*

$$\|\psi_i^{(N)} \circ T_N - \psi_i\|_{L^\infty(\mu)} \lesssim \lambda_i^{m+1} i^{\frac{3}{2}} \sqrt{\frac{\rho_N}{\zeta_N} + \zeta_N\sqrt{\lambda_i}}.$$

Proof We will first modify the $L^2(\mu)$-bound (34) to an $L^2(\mu_N)$-bound using the regularity of eigenfunctions of $-\Delta$, which is then lifted to an $L^\infty(\mu_N)$-bound using Proposition 12 and 14. Finally we use regularity of the eigenfuctions again to transfer the $L^\infty(\mu_N)$-bound to an $L^\infty(\mu)$-bound. To start, we notice that the transport T_N induces a partition of $\mathcal{M}$ by the sets $\{U_i = T_N^{-1}(\{X_i\})\}_{i=1}^N$. Furthermore, the measure preserving property gives $\mu(U_i) = \frac{1}{N}$ and Proposition 6 implies that $U_i \subset B_{\mathcal{M}}(X_i, \rho_N)$. We then have

$$\begin{aligned}\|\psi_i^{(N)} - \psi_i\|_{L^2(\mu_N)}^2 = \frac{1}{N}\sum_{i=1}^N |\psi_i^{(N)}(X_i) - \psi_i(X_i)|^2 &= \sum_{i=1}^N \int_{U_i} |\psi_i^{(N)}(X_i) - \psi_i(X_i)|^2 d\mu(x) \\ &\le \sum_{i=1}^N \int_{U_i} 2|\psi_i^{(N)}(X_i) - \psi_i(x)|^2 + 2|\psi_i(x) - \psi_i(X_i)|^2 d\mu(x),\end{aligned}$$

where

$$\int_{U_i} |\psi_i(x) - \psi_i(X_i)|^2 d\mu(x) \leq \frac{1}{N} \|\nabla \psi_i\|_\infty^2 \rho_N^2.$$

Hence by Proposition 13 and Weyl's law

$$\|\psi_i^{(N)} - \psi_i\|_{L^2(\mu_N)}^2 \leq 2\|\psi_i^{(N)} \circ T_N - \psi_i\|_{L^2(\mu)}^2 + C i^{\frac{m+1}{m}} \rho_N^2.$$

Since $i^{\frac{m}{m+1}} \rho_N^2$ is of higher order than (34), we conclude that

$$\|\psi_i^{(N)} - \psi_i\|_{L^2(\mu_N)}^2 \lesssim i^3 \left(\frac{\rho_N}{\zeta_N} + \zeta_N \sqrt{\lambda_i} \right). \tag{37}$$

Now let $g = \psi_i^{(N)} - \tilde{\psi}_i$ with $\tilde{\psi}_i = \psi_i|_{\{X_1, \ldots, X_N\}}$. We have by Proposition 12

$$\begin{aligned}
\|\Delta_N g\|_{L^\infty(\mu_N)} &\leq \|\Delta_N \psi_i^{(N)} - \Delta \psi_i\|_{L^\infty(\mu_N)} + \|\Delta \psi_i - \Delta_N \tilde{\psi}_i\|_{L^\infty(\mu_N)} \\
&\leq \|\lambda_i^{(N)} \psi_i^{(N)} - \lambda_i \psi_i\|_{L^\infty(\mu_N)} + C(1 + \|\psi_i\|_{\mathcal{C}^3(\mathcal{M})}) \zeta_N \\
&\leq \lambda_i^{(N)} \|\psi_i^{(N)} - \psi_i\|_{L^\infty(\mu_N)} + |\lambda_i^{(N)} - \lambda_i| \|\psi_i\|_{L^\infty(\mu_N)} + C(1 + \|\psi_i\|_{\mathcal{C}^3(\mathcal{M})}) \zeta_N \\
&\leq \lambda_i^{(N)} \|g\|_{L^\infty(\mu_N)} + C \lambda_i^{\frac{m+5}{2}} \left(\frac{\rho_N}{\zeta_N} + \zeta_N \right),
\end{aligned}$$

where we have used (33) and Proposition 13 in the last step. Therefore recalling the definition in (35), we have

$$\lambda_g \leq \lambda_i^{(N)} + C \|g\|_{L^\infty(\mu_N)}^{-1} \lambda_i^{\frac{m+5}{2}} \left(\frac{\rho_N}{\zeta_N} + \zeta_N \right).$$

If $\lambda_g \geq \lambda_i^{(N)} + 1$, then we have

$$\|g\|_{L^\infty(\mu_N)} \leq C \lambda_i^{\frac{m+5}{2}} \left(\frac{\rho_N}{\zeta_N} + \zeta_N \right). \tag{38}$$

Otherwise if $\lambda_g \leq \lambda_i^{(N)} + 1$, then Proposition 14 and (37) implies

$$\|g\|_{L^\infty(\mu_N)} \leq C \left[\lambda_i^{(N)} + 1 \right]^{m+1} \|g\|_{L^1(\mu_N)} \leq C \lambda_i^{m+1} \|g\|_{L^2(\mu_N)} \leq C \lambda_i^{m+1} i^{\frac{3}{2}} \sqrt{\frac{\rho_N}{\zeta_N} + \zeta_N \sqrt{\lambda_i}}, \tag{39}$$

where we have used the fact that $\lambda_i^{(N)} \leq C \lambda_i$ and $\|g\|_{L^1(\mu_N)} \leq \|g\|_{L^2(\mu_N)}$. Comparing (39) with (38), we see that the error in (39) dominates. To finish, we need again to lift the $L^\infty(\mu_N)$-bound to a $L^\infty(\mu)$-bound using regularity of the ψ_i's. Notice that

$$\begin{aligned}
\|\psi_i^{(N)} \circ T_N - \psi_i\|_{L^\infty(\mu)} &= \max_{1 \leq j \leq N} \sup_{x \in U_j} |\psi_i^{(N)} \circ T_N(x) - \psi_i(x)| \\
&\leq \max_{1 \leq j \leq N} \sup_{x \in U_j} \left(|\psi_i^{(N)}(X_j) - \psi_i(X_j)| + |\psi_i(X_j) - \psi_i(x)| \right) \\
&\leq \|\psi_i^{(N)} - \psi_i\|_{L^\infty(\mu_N)} + \|\nabla \psi_i\|_{L^\infty(\mu)} \rho_N \\
&\lesssim \lambda_i^{m+1} i^{\frac{3}{2}} \sqrt{\frac{\rho_N}{\zeta_N} + \zeta_N \sqrt{\lambda_i}},
\end{aligned}$$

where in the last step $\|\nabla\psi_i\|_{L^\infty(\mu)}\rho_N$ is a higher order term that we drop. ■

Now we are finally ready to show the $L^\infty(\mu)$ convergence of $W_n^{\mathcal{M}}$ towards $W^{\mathcal{M}}$.

Proof [Proof of Theorem 7] Recall that

$$W_n^{\mathcal{M}} = \sum_{i=1}^{k_N} \left[1+\lambda_i^{(N)}\right]^{-\frac{s}{2}} \xi_i \psi_i^{(N)} \circ T_N,$$
$$W^{\mathcal{M}} = \sum_{i=1}^{\infty} (1+\lambda_i)^{-\frac{s}{2}} \xi_i \psi_i.$$

Consider two intermediate quantities:

$$\widetilde{W}_n^{\mathcal{M}} = \sum_{i=1}^{k_N} (1+\lambda_i)^{-\frac{s}{2}} \xi_i \psi_i^{(N)} \circ T_N,$$
$$\widehat{W}_n^{\mathcal{M}} = \sum_{i=1}^{k_N} (1+\lambda_i)^{-\frac{s}{2}} \xi_i \psi_i.$$

By the triangle inequality,

$$\begin{aligned}\mathbb{E}_\xi\|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 &\le \mathbb{E}_\xi\Big(\|W_n^{\mathcal{M}} - \widetilde{W}_n^{\mathcal{M}}\|_{L^\infty(\mu)} + \|\widetilde{W}_n^{\mathcal{M}} - \widehat{W}_n^{\mathcal{M}}\|_{L^\infty(\mu)} + \|\widehat{W}_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}\Big)^2 \\ &\le 3\Big(\mathbb{E}_\xi\|W_n^{\mathcal{M}} - \widetilde{W}_n^{\mathcal{M}}\|_{L^\infty(\mu)}^2 + \mathbb{E}_\xi\|\widetilde{W}_n^{\mathcal{M}} - \widehat{W}_n^{\mathcal{M}}\|_{L^\infty(\mu)}^2 + \mathbb{E}_\xi\|\widehat{W}_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2\Big),\end{aligned}$$

so it suffices to bound each term. First, by Proposition 13 and Weyl's law,

$$\begin{aligned}\mathbb{E}_\xi\|\widehat{W}_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 &\le \mathbb{E}_\xi\left[\sum_{i=k_N+1}^{\infty}\sum_{j=k_N+1}^{\infty}(1+\lambda_i)^{-\frac{s}{2}}(1+\lambda_j)^{-\frac{s}{2}}|\xi_i||\xi_j|\|\psi_i\|_{L^\infty(\mu)}\|\psi_j\|_{L^\infty(\mu)}\right] \\ &\lesssim \sum_{i=k_N+1}^{\infty}\sum_{j=k_N+1}^{\infty}(1+\lambda_i)^{-\frac{s}{2}}(1+\lambda_j)^{-\frac{s}{2}}\lambda_i^{\frac{m-1}{4}}\lambda_j^{\frac{m-1}{4}} \\ &= \left[\sum_{i=k_N+1}^{\infty}(1+\lambda_i)^{-\frac{s}{2}}\lambda_i^{\frac{m-1}{4}}\right]^2 \lesssim \left[\int_{k_N}^{\infty} x^{\frac{-2s+m-1}{2m}}dx\right]^2 \lesssim k_N^{\frac{-2s+3m-1}{m}}.\end{aligned} \tag{40}$$

Similarly,

$$\mathbb{E}_\xi\|W_n^{\mathcal{M}} - \widetilde{W}_n^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim \left[\sum_{i=1}^{k_N}\left|\left[1+\lambda_i^{(N)}\right]^{-\frac{s}{2}} - \left[1+\lambda_i\right]^{-\frac{s}{2}}\right| \|\psi_i^{(N)}\circ T_N\|_{L^\infty(\mu)}\right]^2.$$

By Lipschitz continuity of $x^{-s/2}$ on $[1,\infty)$ and (33), for $i = 1, \dots, k_N$,

$$\begin{aligned}\left|\left[1+\lambda_i^{(N)}\right]^{-\frac{s}{2}} - \left[1+\lambda_i\right]^{-\frac{s}{2}}\right| &\le \left[(1+\lambda_i^{(N)}) \wedge (1+\lambda_i)\right]^{-\frac{s}{2}-1} \left|\lambda_i^{(N)} - \lambda_i\right| \\ &\lesssim \left[(1+\lambda_i^{(N)}) \wedge (1+\lambda_i)\right]^{-\frac{s}{2}-1} \lambda_i \left(\frac{\rho_N}{\zeta_N} + \zeta_N\sqrt{\lambda_i}\right) \\ &\lesssim \lambda_i^{-\frac{s}{2}} \left(\frac{\rho_N}{\zeta_N} + \zeta_N\sqrt{\lambda_i}\right).\end{aligned}$$

By (34), if we choose k_N so that

$$k_N^3 \left(\frac{\rho_N}{\zeta_N} + \zeta_N\sqrt{\lambda_{k_N}}\right) \ll 1, \tag{41}$$

then we have, for $i = 1, \dots, k_N$,

$$\|\psi_i^{(N)} \circ T_N\|_{L^\infty(\mu)} \le \|\psi_i^{(N)} \circ T_N - \psi_i\|_{L^\infty(\mu)} + \|\psi_i\|_{L^\infty(\mu)} \lesssim \lambda_i^{\frac{m-1}{4}}.$$

Therefore

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - \widetilde{W}_n^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim \left[\sum_{i=1}^{k_N} \lambda_i^{-\frac{s}{2}} \left(\frac{\rho_N}{\zeta_N} + \zeta_N\sqrt{\lambda_i}\right) \lambda_i^{\frac{m-1}{4}}\right]^2 \lesssim \left(\frac{\rho_N}{\zeta_N} + \zeta_N\sqrt{\lambda_{k_N}}\right)^2, \tag{42}$$

where we have used that $\lambda_i^{\frac{-2s+m-1}{4}}$ is summable for $s > \frac{3}{2}m - \frac{1}{2}$. Lastly, by Lemma 15,

$$\begin{aligned}\mathbb{E}_\xi \|\widetilde{W}_n^{\mathcal{M}} - \widehat{W}_n^{\mathcal{M}}\|_{L^\infty(\mu)}^2 &\lesssim \left[\sum_{i=1}^{k_N} (1+\lambda_i)^{-\frac{s}{2}} \|\psi_i^{(N)} \circ T_N - \psi_i\|_{L^\infty(\mu)}\right]^2 \\ &\lesssim \left[\sum_{i=1}^{k_N} (1+\lambda_i)^{-\frac{s}{2}} \lambda_i^{m+1} i^{\frac{3}{2}} \sqrt{\frac{\rho_N}{\zeta_N} + \zeta_N\sqrt{\lambda_i}}\right]^2 \\ &\lesssim \left[1 \vee k_N^{\frac{-2s+9m+5}{m}}\right] \left(\frac{\rho_N}{\zeta_N} + \zeta_N\right).\end{aligned} \tag{43}$$

Combining (40), (42), (43), we have

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim k_N^{\frac{-2s+3m-1}{m}} + \left[1 \vee k_N^{\frac{-2s+9m+5}{m}}\right] \left(\frac{\rho_N}{\zeta_N} + \zeta_N\right). \tag{44}$$

Now we will set ζ_N and k_N based on the dimension m and the smoothness parameter s.

Case 1: $s \le \frac{9}{2}m + \frac{5}{2}$. Let $\delta > 0$ be arbitrary and set

$$\begin{cases} \zeta_N \asymp N^{-\frac{1}{m+4+\delta}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{1}{(m+4+\delta)(6+\frac{6}{m})}} (\log N)^{-\frac{mp_m}{12m+12}}, & m \le 4, \\ \zeta_N \asymp N^{-\frac{1}{2m}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{1}{12m+12}} (\log N)^{-\frac{mp_m}{12m+12}}, & m \ge 5. \end{cases}$$

One can check that the conditions in (36) and (41) are satisfied and we get

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim \begin{cases} N^{-\frac{2s-3m+1}{(m+4+\delta)(6m+6)}} (\log N)^{\frac{p_m(2s-3m+1)}{12m+12}}, & m \le 4, \\ N^{-\frac{2s-3m+1}{m(12m+12)}} (\log N)^{\frac{p_m(2s-3m+1)}{12m+12}}, & m \ge 5. \end{cases}$$

Case 2: $s > \frac{9}{2}m + \frac{5}{2}$. Now (44) simplifies to

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim k_N^{\frac{-2s+3m-1}{m}} + \left(\frac{\rho_N}{\zeta_N} + \zeta_N\right).$$

Therefore by setting

$$\begin{cases} \zeta_N \asymp N^{-\frac{1}{m+4+\delta}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{m}{(m+4+\delta)(2s-3m+1)}} (\log N)^{-\frac{mp_m}{4s-6m+2}}, & m \le 4, \\ \zeta_N \asymp N^{-\frac{1}{2m}} (\log N)^{\frac{p_m}{2}}, \quad k_N \asymp N^{\frac{1}{4s-6m+2}} (\log N)^{-\frac{mp_m}{4s-6m+2}}, & m \ge 5, \end{cases}$$

we have

$$\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \lesssim \begin{cases} N^{-\frac{1}{m+4+\delta}} (\log N)^{\frac{p_m}{2}}, & m \le 4, \\ N^{-\frac{1}{2m}} (\log N)^{\frac{p_m}{2}}, & m \ge 5. \end{cases}$$

Again one can check that the conditions in (36) and (41) are satisfied. This finishes the proof.

■

4.4 Putting Everything Together

Now we are ready to prove Theorem 1.

Proof [Proof of Theorem 1] By Subsection 4.1, it suffices to show that

$$\mathbb{E}_X \mathbb{E}_{f_0} \Pi_n^{\mathcal{M}} \left(f \in \mathcal{I}(L^\infty(\mu_{N_n})) : \|f - f_0\|_n \ge M_n \varepsilon_n \,|\, \mathcal{X}_{N_n}, \mathcal{Y}_n\right) \xrightarrow{n\to\infty} 0,$$

with the said ε_n. Let A_n be the high probability event in Corollary 8. Denote $F_n(\mathcal{X}_{N_n}, \mathcal{Y}_n) := \Pi_n^{\mathcal{M}}(f \in \mathcal{I}(L^\infty(\mu_{N_n})) : \|f - f_0\|_n \ge M_n \varepsilon_n \,|\, \mathcal{Y}_n)$. We have

$$\mathbb{E}_X \mathbb{E}_{f_0} F_n(\mathcal{X}_{N_n}, \mathcal{Y}_n) = \mathbb{E}_X[\mathbb{E}_{f_0} F_n(\mathcal{X}_{N_n}, \mathcal{Y}_n)]\mathbf{1}_{A_n} + \mathbb{E}_X[\mathbb{E}_{f_0} F_n(\mathcal{X}_{N_n}, \mathcal{Y}_n)]\mathbf{1}_{A_n^c}.$$

Since $F_n(\mathcal{X}_{N_n}, \mathcal{Y}_n) \le 1$, the second term is upper bounded by $\mathbb{P}_X(A_n^c) \to 0$. It then suffices to show $\mathbb{E}_{f_0} F_n(\mathcal{X}_{N_n}, \mathcal{Y}_n) \to 0$ in the event of A_n. By Corollary 8 we can have $10\mathbb{E}_\xi \|W_n^{\mathcal{M}} - W^{\mathcal{M}}\|_{L^\infty(\mu)}^2 \le n^{-1}$ if we set a large enough proportion constant for N_n. This together with Theorem 5 and (van der Vaart and van Zanten, 2008a, Theorem 2.2) implies that there exists sets $B_n \subset \mathcal{I}(L^\infty(\mu_{N_n}))$ (by taking the intersection of the sets B_n in Theorem 5 and $\mathcal{I}(L^\infty(\mu_{N_n}))$) so that, for the same ε_n in the theorem statement,

$$\begin{aligned} \log N(6\varepsilon_n, B_n, \|\cdot\|_{L^\infty(\mu)}) &\le 24Cn\varepsilon_n^2, \\ \pi_n^{\mathcal{M}}(B_n^c) &\le e^{-4Cn\varepsilon_n^2}, \\ \pi_n^{\mathcal{M}}(\|w - \Phi^{-1}(f_0)\|_{L^\infty(\mu)} < 4\varepsilon_n) &\ge e^{-4n\varepsilon_n^2}, \end{aligned}$$

where $\pi_n^{\mathcal{M}} = \mathcal{L}(W_n^{\mathcal{M}})$.

Case 1: Regression. In the regression case, since Φ is the identity, we have $\Pi_n^{\mathcal{M}} = \pi_n^{\mathcal{M}}$ and the above three conditions are true with $\pi_n^{\mathcal{M}}$ replaced by $\Pi_n^{\mathcal{M}}$. Furthermore, since $\|\cdot\|_n$ is upper bounded by $\|\cdot\|_{L^\infty(\mu)}$, the above conditions remain true for the empirical norm. This together with the general results in (Ghosal and van der Vaart, 2017, Section 8.3.2) proves the assertion.

Case 2: Classification. In the classification setting, by (Ghosal and van der Vaart, 2017, Lemma 2.8) the average Kullback-Leibler divergence and variance in (17) between the densities after applying the link function Φ are upper bounded by a multiple of the empirical norm. In particular if we set $\mathcal{B}_n = \Phi(B_n)$, then the conditions in (14), (15) and (16) hold, for a possibly different proportion constant for ε_n. Moreover, as discussed before, the root average square Hellinger distance d_n in (19) is equivalent to the empirical norm. Hence the result follows by (Ghosal and van der Vaart, 2007, Theorem 4). ■

5. Discussion

In this paper we have analyzed graph-based SSL using recent results on spectral convergence of graph Laplacians and standard Bayesian nonparametrics techniques. We show that, for a suitable choice of prior constructed with sufficiently many unlabeled data, the posterior contracts around the truth at a rate that is minimax optimal up to logarithmic factor. Our theory applies to both regression and classification.

We have assumed throughout that the X_i's are uniformly distributed on $\mathcal{M}$. Our results can be generalized to nonuniform positive density q with respect to the volume form. In such a case, the continuum field is the Gaussian measure $\mathcal{N}(0, (I - \Delta_q))^{-s}$ where $-\Delta_q := -\frac{1}{q}\mathrm{div}(q^2\nabla)$ is a weighted Laplacian-Beltrami operator, for which spectral convergence results can be found in García Trillos et al. (2019a). Since we do not have an explicit dependence of the proportion constant on the index i as in Proposition 10, we have chosen to present our result in the uniform case. However, similar conclusions should be expected to hold in the general case.

The Bayesian methodology we analyzed is inspired by popular existing graph-based optimization methods for SSL (Belkin et al., 2004; Zhu, 2005). In order to achieve optimal contraction, the prior smoothness parameter needs to match the regularity of the truth, which is rarely available in applications. An important research direction stemming from our work is the development and analysis of adaptive Bayesian SSL methodologies that can achieve optimal contraction without a priori smoothness information. We expect that the existing results on adaptive estimation on manifolds (Castillo et al., 2014) and on large graphs (Kirichenko and van Zanten, 2017) will be an important stepping stone in this direction.

Acknowledgment

Both authors are thankful for the support of NSF and NGA through the grant DMS-2027056. The work of DSA was also partially supported by the NSF Grant DMS-1912818/1912802.

References

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1-3):209–239, 2004.

Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *International Conference on Computational Learning Theory*, pages 624–638. Springer, 2004.

Andrea L Bertozzi, Bamdad Hosseini, Hao Li, Kevin Miller, and Andrew M Stuart. Posterior consistency of semi-supervised regression on graphs. *Inverse Problems*, 37(10):105011, 2021.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*, pages 177–186. Institute of Mathematical Statistics, 2007.

Jeff Calder and Nicolás García Trillos. Improved spectral convergence rates for graph Laplacians on ε-graphs and k-nn graphs. *Applied and Computational Harmonic Analysis*, 2022.

Jeff Calder, Dejan Slepčev, and Matthew Thorpe. Rates of convergence for Laplacian semi-supervised learning with low labeling rates. *arXiv preprint arXiv:2006.02765*, 2020.

Jeff Calder, Nicolás García Trillos, and Marta Lewicka. Lipschitz regularity of graph Laplacians on random data clouds. *SIAM Journal on Mathematical Analysis*, 54(1):1169–1222, 2022.

Yaiza Canzani. Analysis on manifolds via the Laplacian. *Lecture Notes available at: http://www. math. harvard. edu/canzani/docs/Laplacian. pdf*, 2013.

Ismaël Castillo, Gérard Kerkyacharian, and Dominique Picard. Thomas Bayes' walk on manifolds. *Probability Theory and Related Fields*, 158(3):665–710, 2014.

Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised Learning. *Cambridge, Massachusettes: The MIT Press View Article*, 2, 2006.

Jose A Costa and Alfred O Hero. Determining intrinsic dimension and entropy of high-dimensional shape spaces. In *Statistics and Analysis of Shapes*, pages 231–252. Springer, 2006.

Thierry Coulhon, Gerard Kerkyacharian, and Pencho Petrushev. Heat kernel generated frames in the setting of Dirichlet spaces. *Journal of Fourier Analysis and Applications*, 18(5):995–1066, 2012.

Fábio Gagliardi Cozman and Ira Cohen. Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. *Semi-supervised Learning*, 4:57–72, 2006.

Harold Donnelly. Eigenfunctions of the Laplacian on compact Riemannian manifolds. *Asian Journal of Mathematics*, 10(1):115–126, 2006.

David B Dunson, Hau-Tieng Wu, and Nan Wu. Spectral convergence of graph Laplacian and heat kernel reconstruction in L^∞ from random samples. *Applied and Computational Harmonic Analysis*, 55:282–336, 2021.

David Eric Edmunds and Hans Triebel. *Function Spaces, Entropy Numbers, Differential Operators*, volume 120. Cambridge University Press Cambridge, 1996.

Ahmed El Alaoui, Xiang Cheng, Aaditya Ramdas, Martin J Wainwright, and Michael I Jordan. Asymptotic behavior of L_p-based Laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

Nicolás García Trillos and D. Sanz-Alonso. Continuum limits of posteriors in graph Bayesian inverse problems. *SIAM Journal on Mathematical Analysis*, 50(4):4020–4040, 2018.

Nicolás García Trillos, Moritz Gerlach, Matthias Hein, and Dejan Slepčev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator. *Foundations of Computational Mathematics*, pages 1–61, 2019a.

Nicolás García Trillos, Daniel Sanz-Alonso, and Ruiyi Yang. Local regularization of noisy point clouds: Improved global geometric estimates and data analysis. *Journal of Machine Learning Research*, 20(136):1–37, 2019b.

Nicolás García Trillos, Z. Kaplan, T. Samakhoana, and D. Sanz-Alonso. On the consistency of graph-based Bayesian semi-supervised learning and the scalability of sampling algorithms. *Journal of Machine Learning Research*, 21(28):1–47, 2020.

Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press, 2017.

Subhashis Ghosal, Jayanta K Ghosh, and Aad van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, pages 500–531, 2000.

John Harlim, Daniel Sanz-Alonso, and Ruiyi Yang. Kernel methods for Bayesian elliptic inverse problems on manifolds. *SIAM/ASA Journal on Uncertainty Quantification*, 8(4): 1414–1445, 2020.

Matthias Hein and Jean-Yves Audibert. Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 289–296, 2005.

Franca Hoffmann, Bamdad Hosseini, Zhi Ren, and Andrew M Stuart. Consistency of semi-supervised learning algorithms on graphs: Probit and one-hot methods. *Journal of Machine Learning Research*, 21:1–55, 2020.

Alisa Kirichenko and Harry van Zanten. Estimating a smooth function on a large graph by Bayesian Laplacian regularisation. *Electronic Journal of Statistics*, 11(1):891–915, 2017.

Alexander Kushpel and Jeremy Levesley. Entropy of Sobolev's classes on compact homogeneous Riemannian manifolds. *arXiv preprint arXiv:1504.06508*, 2015.

Annika Lang, Jürgen Potthoff, Martin Schlather, and Dimitri Schwab. Continuity of random fields on Riemannian manifolds. *Communications on Stochastic Analysis*, 10(2):4, 2016.

Wenbo V Li and Werner Linde. Approximation, metric entropy and small ball estimates for Gaussian measures. *The Annals of Probability*, 27(3):1556–1578, 1999.

Feng Liang, Sayan Mukherjee, and Mike West. The use of unlabeled data in predictive modeling. *Statistical Science*, pages 189–205, 2007.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (4):423–498, 2011.

Boaz Nadler, Nathan Srebro, and Xueyuan Zhou. Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data. *Advances in Neural Information Processing Systems*, 22:1330–1338, 2009.

Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14(5), 2013.

Yiling Qiao, Chang Shi, Chenjian Wang, Hao Li, Matt Haberland, Xiyang Luo, Andrew M Stuart, and Andrea L Bertozzi. Uncertainty quantification for semi-supervised multi-class classification in image processing and ego-motion analysis of body-worn videos. *Electronic Imaging*, 2019(11):264–1, 2019.

Daniel Sanz-Alonso and Ruiyi Yang. The SPDE approach to Matérn fields: Graph representations. *To appear in Statistical Science*, 2022a.

Daniel Sanz-Alonso and Ruiyi Yang. Finite element representations of Gaussian processes: Balancing numerical and statistical accuracy. *To appear in SIAM/ASA Journal on Uncertainty Quantification*, 2022b.

Matthias Seeger. Learning with labeled and unlabeled data. Technical report, 2000.

Xiaotong Shen and Larry Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29(3):687–714, 2001.

Lei Shi, Rada Mihalcea, and Mingjun Tian. Cross language text classification by model translation and semi-supervised learning. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1057–1067, 2010.

Aarti Singh, Robert Nowak, and Jerry Zhu. Unlabeled data: Now it helps, now it doesn't. *Advances in Neural Information Processing Systems*, 21, 2008.

John A Toth and Steve Zelditch. Riemannian manifolds with uniformly bounded eigenfunctions. *Duke Mathematical Journal*, 111(1):97–132, 2002.

Hans Triebel. *Theory of Function Spaces II*. Monographs in Mathematics. Springer Basel, 1992.

Aad van der Vaart and Harry van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008a.

Aad van der Vaart and Harry van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pages 200–222. Institute of Mathematical Statistics, 2008b.

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

Larry Wasserman and John D Lafferty. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems*, pages 801–808, 2008.

Jason Weston, Christina Leslie, Eugene Ie, Dengyong Zhou, Andre Elisseeff, and William Stafford Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, 2005.

Bin Xu. Asymptotic behavior of L^2-normalized eigenfunctions of the Laplace-Beltrami operator on a closed Riemannian manifold. *Harmonic Analysis and its Applications*, pages 99–117, 2006.

Yun Yang and David B Dunson. Bayesian manifold regression. *The Annals of Statistics*, 44(2):876–905, 2016.

Xiaojin Zhu. *Semi-supervised learning with graphs*. Carnegie Mellon University, 2005.