Journal of Machine Learning Research 21 (2020) 1-47 Submitted 11/17; Revised 9/19; Published 03/20

On the consistency of graph-based Bayesian semi-supervised learning and the scalability of sampling algorithms

Nicolás García Trillos GARCIATRILLO@WISC.EDU
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706, USA

Zachary Kaplan ZACHARY.ABRAHAM.KAPLAN@GMAIL.COM
Division of Applied Mathematics
Brown University
Providence, RI 02912, USA

Thabo Samakhoana THABO_SAMAKHOANA@ALUMNI.BROWN.EDU
Division of Applied Mathematics
Brown University
Providence, RI 02912, USA

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

Editor: Sanjoy Dasgupta

Abstract

This paper considers a Bayesian approach to graph-based semi-supervised learning. We show that if the graph parameters are suitably scaled, the graph-posteriors converge to a continuum limit as the size of the unlabeled data set grows. This consistency result has profound algorithmic implications: we prove that when consistency holds, carefully designed Markov chain Monte Carlo algorithms have a uniform spectral gap, independent of the number of unlabeled inputs. Numerical experiments illustrate and complement the theory.

Keywords: semi-supervised learning, graph-based learning, Markov chain Monte Carlo, spectral gap

1. Introduction

The aim of this paper is to contribute to the theoretical and methodological understanding of graph-based semi-supervised learning and its Bayesian formulation. Semi-supervised learning makes use of *labeled* and *unlabeled* data for training. Labeled data consists of pairs of inputs and outputs, while unlabeled data consists only of inputs. We focus on the inductive learning task of inferring the hidden map from inputs to outputs. We work under the classical assumption that the inputs are concentrated on a low dimensional manifold embedded in a higher dimensional ambient space. Traditional graph-based optimization methods find

†. All authors contributed equally to this work.

a suitable input/output map by minimizing an objective functional comprising of at least two terms:

i) A regularization term involving a graph-Laplacian built using only the input data. Regularization promotes smoothness of the recovered map along the input manifold.

ii) A data-misfit term that promotes that the recovered map is accurate over the labeled data.

Graph-based optimization methods will be reviewed below. In this paper we study a graph-based *Bayesian* approach that, instead of recovering a single input/output map, gives a *posterior* probability distribution over maps. The posterior contains information on the most likely maps to have produced the training data, but also on the uncertainty remaining in the recovery. As in optimization methods, the Bayesian posterior is found by balancing a smoothness penalty and a data misfit penalty. These competing forces are encoded in a prior distribution and a likelihood function.

I) The prior distribution serves as a regularization that promotes maps that satisfy certain smoothness conditions. The prior covariance will be defined using a graph-Laplacian built using only the input data.

II) The likelihood function plays the role of a data-misfit functional and promotes maps that are accurate on the labeled data.

We investigate the convergence of posterior distributions and the scaling of sampling algorithms in the limit of training large numbers of unlabeled examples. We consider ε-graphs, which connect any two inputs whose distance is less than ε. Our results guarantee that, provided that the connectivity parameter ε is suitably scaled with the number of inputs, the graph-based posteriors converge, as the size of the unlabeled data set grows, to a continuum posterior. Moreover we show that, under the existence of a continuum limit, carefully designed graph-based Markov chain Monte Carlo (MCMC) sampling algorithms have a uniform spectral gap, independent of the number of unlabeled examples. Roughly speaking our results imply that the number of Markov chain iterations needed to achieve a given accuracy is independent of the number of unlabeled data points. However, the cost per iteration will, in general, depend on the size of the data-set.

The continuum limit theory that we bring forward is of interest in three distinct ways. First, it establishes the statistical consistency of graph-based semi-supervised learning methods in machine learning; second, it suggests suitable scalings of graph parameters of practical interest (e.g. see the conditions in the parameter s in Theorem 3, the scalings for the graph connectivity ε also in Theorem 3, and the truncation point for the spectrum of the graph-Laplacian in (15) used to construct the prior in Theorem 3); and third, statistical consistency is shown to go hand in hand with algorithmic scalability: when graph-based learning problems have a continuum limit, algorithms that exploit this limit structure converge in a number of iterations that is independent of the size of the unlabeled data set. The theoretical understanding of these questions relies heavily on recently developed bounds for the asymptotic behavior of the spectra of graph-Laplacians.

Our presentation brings together various approaches to semi-supervised learning, and highlights the similarities and differences between optimization and Bayesian formulations. We include a computational study that suggests directions for further theoretical developments, and illustrates the non-asymptotic relevance of our asymptotic results.

1.1. Problem Description

We now provide a brief intuitive problem description; a fully rigorous account is given in section 2. We highlight the generality of our setting, which covers a wide class of methods for semi-supervised regression and classification, including probit and logistic Bayesian methods.

We assume to be given n inputs lying on an *unknown* m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$, p of which are labeled. The collection of input data will be denoted by $\mathcal{M}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathcal{M}$, and we denote by $y \in \mathbb{R}^p$ the vector of labels. The pairs of inputs/outputs $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_p, y_p)$ form the labeled data and the inputs $\mathbf{x}_{p+1}, \ldots, \mathbf{x}_n$ are unlabeled examples. Our goal is to use the observed data to learn a label for each point in the input space (assumed to be the unknown manifold $\mathcal{M}$).

In the ideal case of known manifold $\mathcal{M}$, a standard Bayesian approach to such learning task proceeds by putting a Gaussian process prior $\boldsymbol{\pi} = N(0, -\Delta_{\mathcal{M}}^s)$ over mappings $u : \mathcal{M} \to \mathbb{R}$ and proposing a statistical model (e.g. additive Gaussian noise, probit, logistic) for the data which is encoded in a negative log-likelihood Φ . The data model may depend on a forward map $\mathcal{F}$ that first transforms the input/output function u, and on the subsequent application of an observation map $\mathcal{O}$; see section 2.1.2 for concrete choices of forward and observation maps considered in this paper. In the above, $\Delta_{\mathcal{M}}$ denotes the Laplace Beltrami operator on $\mathcal{M}$ and the parameter $s > 0$ determines the regularity of prior draws; more intuition on the role of the Laplace Beltrami operator and the parameter s will be given below. Combining the prior and the likelihood via Bayes' rule, one can define a posterior distribution $\boldsymbol{\mu}$ over functions $u : \mathcal{M} \to \mathbb{R}$ by

$$\boldsymbol{\mu}(du) \propto \exp\bigl(-\Phi(u; y)\bigr)\, \boldsymbol{\pi}(du). \tag{1}$$

That is, the posterior is the distribution whose density with respect to the prior is proportional to the likelihoood function.

However, as the input space $\mathcal{M}$ is assumed to be unknown, the above Bayesian formulation is impractical. We follow instead an *intrinsic* approach and aim first at finding suitable labels for the inputs in the given point cloud $\mathcal{M}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, which are then extrapolated, via a Voronoi extension (or 1-NN extension), to assign a label to every point on the manifold $\mathcal{M}$ or in the ambient space $\mathbb{R}^d$. We take a Bayesian approach to learn a discrete input/output function $u_n : \mathcal{M}_n \to \mathbb{R}$ by first building a graph Laplacian which induces a Gaussian prior distribution $\boldsymbol{\pi_n} = N(0, \Delta_{\mathcal{M}_n}^{-s})$ over discrete functions u_n, and then introducing an approximatoin Φ_n to the negative log-likelihood function Φ. In this way, geometric properties of the underlying manifold $\mathcal{M}$ are extracted from the point cloud $\mathcal{M}_n$ and incorporated both in the prior and the likelihood. Notice that if the original data model is defined in terms of some forward and observation maps, then one should also construct appropriate graph approximations for them (see section 2.2.3 for the approximation of the forward and observation maps considered in section 2.1.2). The solution of the graph-based

Bayesian approach is a posterior distribution over discrete functions

$$\boldsymbol{\mu_n}(du_n) \propto \exp\bigl(-\Phi_n(u_n; y)\bigr)\, \boldsymbol{\pi_n}(du_n). \tag{2}$$

The details on how we construct —without use of the ambient space or $\mathcal{M}$— the graph-based prior $\boldsymbol{\pi_n}$ and likelihood Φ_n are given in section 2.

Two interpretations of equations (1) and (2) will be useful. The first one is to see (2) as a graph-based discretization of a Bayesian inverse problem over functions on $\mathcal{M}$ whose posterior solution is given by equation (1). The second is to interpret them as classical Bayesian regression problems. In the latter interpretation, $\mathcal{M}$ may represent a low-dimensional manifold sufficient to characterize features living in an extremely high dimensional ambient space ($m \ll d$), perhaps upon some dimensionality reduction of the given inputs; in the former, $\mathcal{M}$ may represent the unknown physical domain of a differential equation. We note again that our framework covers —by the flexibility in the choice of misfit functional Φ— a wide class of classification and regression learning problems that includes Bayesian probit and logistic models.

Our first goal is to study the large n limit of the posterior distribution $\boldsymbol{\mu_n}$ after it has been pushed-forward by the interpolation map $\mathcal{I}_n^1$ (see definition (4)) that extends functions defined on $\mathcal{M}_n$ to functions defined on $\mathcal{M}$. Our second goal is to study the algorithmic scalability of carefully designed MCMC schemes to sample from $\boldsymbol{\mu_n}$ (see Algorithm 2). The theory on statistical consistency and algorithmic scalability that we set forth concerns regimes with large number n of input training data and moderate number p of labeled examples. This is precisely the regime of interest in semi-supervised learning applications, where often labeled data is expensive to collect but unlabeled data abounds. Our consistency results guarantee that graph-based posteriors of the form (2) are close to a ground truth posterior of the form (1), while the algorithmic scalability that we establish ensures the convergence, in an n-independent number of iterations, of certain MCMC methods for graph posterior sampling. The computational cost per iteration may, however, grow with n. These MCMC methods are in principle applicable in fully supervised learning, but their performance would typically deteriorate if both n and p are allowed to grow. Finally, we note that although our exposition is focused on semi-supervised regression, our conclusions are equally relevant for semi-supervised classification.

1.2. Literature

Here we put into perspective our framework by contrasting it with optimization and extrinsic approaches to semi-supervised learning, and by relating it to other surrogate and approximate methods for Bayesian inversion. We also give some background on MCMC algorithms.

1.2.1. Graph-Based Semi-supervised Learning

We refer to Zhu (2005) for an introductory tutorial on semi-supervised learning with useful pointers to the literature. The question of when and how unlabeled data matters is addressed in Liang et al. (2007). Some key papers on graph-based methods are Zhu et al. (2003); Hartog and van Zanten (2016); Blum and Chawla (2001).

As already noted, a key motivation for graph-based semi-supervised learning is that high dimensional inputs can often be represented in a low-dimensional manifold, whose local geometry may be learned by imposing a graph structure on the inputs. In practice, features may be close to but not exactly *on* an underlying manifold (García Trillos et al., 2019). The question of how to find suitable manifold representations has led to a vast literature on dimensionality reduction techniques and manifold learning, e.g. Roweis and Saul (2000); Tenenbaum et al. (2000); Donoho and Grimes (2003); Belkin and Niyogi (2004).

The reconstruction of the hidden input/output maps from few labeled examples can be carried out by compromising between data fidelity and regularization (along the underlying manifold). Our work considers regularizations defined in terms of the graph Laplacian $\Delta_{\mathcal{M}_n}^{-s}$, with the power parameter $s > 0$ tuning the amount of regularization (the higher s the more regularity imposed). Although the use of such parameter is standard in the machine learning literature (Sindhwani et al., 2005) our work provides new understanding on how s should be chosen in terms of the intrinsic dimension m of the input manifold in order to have consistent learning in the limit of large numbers of unlabeled examples. Our analysis builds on recent results from García Trillos et al. (2018) where explicit rates of convergence for the spectra of graph Laplacians towards the spectrum of a continuum differential operator have been obtained. These results relate in a quantitative way the geometry of the underlying manifold $\mathcal{M}$ and that of the point cloud $\mathcal{M}_n$. The problem of studying the large sample limit of graph Laplacians has received much attention in the last decades. Initially, most results were of pointwise type as in Hein et al. (2007); **?**); Giné and Koltchinskii (2006); Hein (2006); Singer (2006); Ting et al. (2010). More recently, the focus has been given to variational and spectral convergence Belkin and Niyogi (2007); Singer and Wu (2017); García Trillos and Slepčev (2016b); Shi (2015); Burago et al. (2014); García Trillos et al. (2018, 2019).

Alternative graph p-Laplacian regularizations were introduced in Zhou and Schölkopf (2005). This type of regularization is similar to the one considered in this paper, but it does not induce a Gaussian prior on the hidden input/output map; because of this, it is more difficult to implement algorithms to sample from posteriors based on p-Laplacian regularization. The statistical consistency of semi-supervised learning based on p-Laplacian regularization has been studied in El Alaoui et al. (2016), Slepčev and Thorpe (2017). These papers have rigorously analyzed how the parameter p —which plays an analogous role to s in our context— should be chosen in terms of dimension so that “labels are not forgotten" in the large data limit.

1.2.2. Bayesian vs. Optimization, and Intrinsic vs. Extrinsic

In this subsection we focus on the regression interpretation, with labels directly obtained from noisy observation of the unknown input/output function. The Bayesian formulation that we consider has the advantage over traditional optimization formulations in that it allows for uncertainty quantification in the recovery of the unknown function Bertozzi et al. (2018). Moreover, from a computational viewpoint, we shall show that certain sampling algorithms have desirable scaling properties —these algorithms, in the form of simulated annealing, may also find application within optimization formulations (Geyer and Thompson, 1995).

The Bayesian update (2) is intimately related to the optimization problem

$$\min_{u_n} \langle \Delta^s_{\mathcal{M}_n} u_n, u_n \rangle + \Phi_n(u_n; y). \tag{3}$$

Here $\Delta_{\mathcal{M}_n}$ represents the graph-Laplacian, as defined in equation (13) below, and the minimum is taken over square integrable functions on the point cloud $\mathcal{M}_n$. Precisely, the solution u_n^* to (3) is the mode (or MAP for *maximum a posteriori*) of the posterior distribution $\boldsymbol{\mu_n}$ in (2) with a Gaussian prior $\boldsymbol{\pi_n} = N(0, \Delta^{-s}_{\mathcal{M}_n})$.

The Bayesian problem (2) and the variational problem (3) are *intrinsic* in the sense that they are constructed without reference to the ambient space (other than through its metric), working in the point cloud $\mathcal{M}_n$. In order to address the *generalization problem* of assigning labels to points $\mathbf{x} \notin \mathcal{M}_n$ we use interpolation maps that turn functions defined on the point cloud into functions defined on the ambient space. We will restrict our attention to the family of k-NN interpolation maps defined by

$$\big[\mathcal{I}_n^k(u_n)\big](\mathbf{x}) := \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} u_n(\mathbf{x}_i), \quad \mathbf{x} \in \mathbb{R}^d, \tag{4}$$

where $N_k(\mathbf{x})$ is the set of k-nearest neighbors in $\mathcal{M}_n$ to $\mathbf{x}$; here the distance used to define nearest neighbors is that of the ambient space. Within our Bayesian setting we consider $\mathcal{I}_n \sharp \boldsymbol{\mu_n}$, the push-forward of $\boldsymbol{\mu_n}$ by $\mathcal{I}_n$, as the fundamental object that allows us to assign labels to inputs $\mathbf{x} \notin \mathcal{M}_n$, and quantify the uncertainty in such inference. The need of interpolation maps also appears in the context of intrinsic variational approaches to binary classification (García Trillos and Murray, 2017) and in the context of variational problems of the form (3): the function u_n^* is only defined on $\mathcal{M}_n$, and hence should be extended to the ambient space via an interpolation map $\mathcal{I}_n$.

Intrinsic approaches contrast with *extrinsic* ones, such as manifold regularization (Belkin and Niyogi, 2005; Belkin et al., 2006). This method solves a variational problem of the form

$$\min_{u} \langle \Delta^s_{\mathcal{M}_n} u|_{\mathcal{M}_n}, u|_{\mathcal{M}_n} \rangle + \Phi(u; y) + \zeta \|u\|^2_{\mathcal{H}_K}, \tag{5}$$

where now the minimum is taken over functions in a reproducing kernel Hilbert space $\mathcal{H}_K$ defined over the *ambient* space $\mathbb{R}^d$, and $u|_{\mathcal{M}_n}$ denotes the restriction of u to $\mathcal{M}_n$. The kernel K is defined in $\mathbb{R}^d$ and the last term in the objective functional, not present in (3), serves as a regularizer in the ambient space; the parameter $\zeta \geq 0$ controls the weight given to this new term. Bayesian and extrinsic formulations may be combined in future work.

In short, *extrinsic* variational approaches solve a problem of the form (5), and *intrinsic* ones solve (3) and then generalize by using an appropriate interpolation map. In the spirit of the latter, the *intrinsic* Bayesian approach of this paper defines an intrinsic graph-posterior by (2) and then this posterior is pushed-forward by an interpolation map. What are the advantages and disadvantages of each approach? Intuitively, the intrinsic approach seems more natural for label inference of inputs *on* or *close to* the underlying manifold $\mathcal{M}$. However, the extrinsic approach is appealing for problems where no low-dimensional manifold structure is present in the input space.

1.2.3. Approximate and Surrogate Bayesian Learning

Our learning problem can be seen as approximating a ground-truth Bayesian inverse problem over functions on the underlying manifold $\mathcal{M}$ (Dashti and Stuart; García Trillos and Sanz-Alonso, 2017; Harlim et al., 2019). Traditional problem formulations and sampling algorithms require repeated evaluation of the likelihood, often making naive implementations impractical. For this reason, there has been recent interest in reduced order models (Sacks et al., 1989; Kennedy and O'Hagan, 2001; Arridge et al., 2006; Cui et al., 2015), and in defining surrogate likelihoods in terms of Gaussian processes (Rasmussen and Williams, 2006; Stein, 2012; Stuart and Teckentrup, 2017), or polynomial chaos expansions (Xiu, 2010; Marzouk et al., 2007). Pseudo-marginal (**?**) and approximate Bayesian computation methods (Beaumont et al., 2002) have become popular in intractable problems where evaluation of the likelihood is not possible. There are two distinctive aspects of the graph-based models employed here. First, they approximate both the prior and the likelihood; and second, the approximate and ground-truth posteriors live in different spaces: the former is a measure over functions on a point cloud, while the latter is a measure over functions on the continuum. The paper García Trillos and Sanz-Alonso (2018a) studied the continuum limits of graph-posteriors to the ground-truth continuum posterior. This was achieved by using a new topology inspired by the analysis of functionals over functions in point clouds arising in machine learning (García Trillos and Slepčev, 2016a, 2014, 2016b; Slepčev and Thorpe, 2017).

In this paper, we rigorously make a connection between the graph Bayesian model and the continuum one, by proving that in the large number of unlabeled data limit, the extended graph posterior converges towards the posterior of the continuum Bayesian model.

1.2.4. Markov Chain Monte Carlo

MCMC is a popular class of algorithms for posterior sampling. Here we consider certain Metropolis–Hastings MCMC methods that construct a Markov chain that has the posterior as its invariant distribution by sampling from a user-chosen proposal and accepting/rejecting the samples using a general recipe. Posterior expectations are then approximated by averages with respect to the chain's empirical measure. The generality of Metropolis–Hastings algorithms is a double-edged sword: the choice of proposal may have a dramatic impact on the convergence of the chain. Even for a given form of proposal, parameter tuning is often problematic. These issues are exacerbated in learning problems over functions, as traditional algorithms often break-down.

The preconditioned Crank-Nicolson (pCN) algorithm introduced in Beskos et al. (2008) allows for scalable sampling of infinite dimensional functions provided that the target is suitably defined as a change of measure. Indeed, the key idea of the method is to exploit this change of measure structure, that arises naturally in Bayesian nonparameterics but also in the sampling of conditioned diffusions and elsewhere. Robustness is understood in the sense that, when pCN is used to sample functions projected onto a finite D-dimensional space, the rate of convergence of the chain is independent of D. This was already observed in Beskos et al. (2008) and Cotter et al. (2013), and was further understood in Hairer et al. (2014) by showing that projected pCN methods have a uniform spectral gap, while traditional random walk does not.

In this paper we substantiate the use of graph-based pCN MCMC algorithms (Bertozzi et al., 2018) in semi-supervised learning. The main insight is that our continuum limit results provide the necessary change of measure structure for the robustness of pCN. This allows us to establish their uniform spectral gap in the regime where the continuum limit holds. Namely, we show that if the number p of labeled data is fixed, then the rate of convergence of graph pCN methods for sampling graph posterior distributions is independent of n. We remark that pCN addresses some of the challenges arising from sampling functions, but fails to address challenges arising from tall data. Some techniques to address this complementary difficulty are reviewed in Bardenet et al. (2017).

1.3. Paper Organization and Main Contributions

A thorough description of our setting is given in section 2. Algorithms for posterior sampling are presented in section 3. Section 4 contains our main theorems on continuum limits of graph posteriors and uniform spectral gaps. Finally, a computational study is conducted in section 5. All proofs and technical material are collected in an appendix.

The two main theoretical contributions of this paper are Theorem 3 —establishing statistical consistency of intrinsic graph methods generalized by means of interpolation maps— and Theorem 7 —establishing the uniform spectral gap for graph-based pCN methods under the conditions required for the existence of a continuum limit. Both results require appropriate scalings of the graph connectivity with the number of inputs. An important contribution of this paper is the analysis of truncated graph-priors that retain only the portion of the spectra of the graph Laplacian that provably approximates that of the ground-truth continuum. As it turns out, only a portion of the spectrum of the graph Laplacian contains relevant information about the underlying manifold $\mathcal{M}$, and thus one can disregard higher modes. See the discussion in section 5.1.1 and Figure 2 for an illustration of this.

From a numerical viewpoint, our experiments illustrate parameter choices that lead to successful graph-based inversion, highlight the need for a theoretical understanding of the spectrum of graph Laplacians and of regularity of functions on graphs, and show that the asymptotic consistency and scalability analysis set forth in this paper is of practical use outside the asymptotic regime.

2. Setting

Throughout, $\mathcal{M}$ will denote an m-dimensional, compact, smooth manifold embedded in $\mathbb{R}^d$. We let $\mathcal{M}_n := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a collection of i.i.d. samples from the uniform distribution on $\mathcal{M}$. We are interested in learning functions defined on $\mathcal{M}_n$ by using the inputs $\mathbf{x}_i$ and some output values, obtained by noisy evaluation at $p \leq n$ inputs of a transformation of the unknown function. The learning problem in the discrete space $\mathcal{M}_n$ is defined by means of a graph-based discretization of a continuum learning problem defined over functions on $\mathcal{M}$. We view the continuum problem as a ground-truth case where full geometric information of the input space is available. We describe the continuum learning setting in subsection 2.1, followed by the discrete learning setting in subsection 2.2. We will denote by $L^2(\gamma)$ the space of functions on the underlying manifold that are square integrable with respect to the uniform measure γ. We use extensively that functions in $L^2(\gamma)$ can be written in terms of the (normalized) eigenfunctions $\{\psi_i\}_{i=1}^{\infty}$ of the Laplace Beltrami operator $\Delta_{\mathcal{M}}$. We denote by $\{\lambda_i\}_{i=1}^{\infty}$ the associated eigenvalues of $-\Delta_{\mathcal{M}}$, assumed to be in non-decreasing order and repeated according to multiplicity. Analogous notations will be used in the graph-based setting, with scripts n.

2.1. Continuum Learning Setting

Our ground-truth continuum learning problem consists of the recovery of a function $u \in L^2(\gamma)$ from data $y \in \mathbb{R}^p$. The data y are assumed to be a noisy observation of a vector $v \in \mathbb{R}^p$ obtained indirectly from the function of interest u as follows:

$$u \in L^2(\gamma) \mapsto v := \mathcal{O} \circ \mathcal{F}(u) \mapsto y.$$

Here $\mathcal{F} : L^2(\gamma) \to L^2(\gamma)$ is interpreted as a *forward map* representing, for instance, a map from inputs to outputs of a differential equation. As a particular case of interest, $\mathcal{F}$ may be the identity map in $L^2(\gamma)$. The map $\mathcal{O} : L^2(\gamma) \to \mathbb{R}^p$ is interpreted as an observation map, and is assumed to be linear and continuous. The Bayesian approach that we will now describe proceeds by specifying a prior on the unknown function u, and a noise model for the generation of data y given the vector $v = \mathcal{O} \circ \mathcal{F}(u)$. The solution is a posterior measure $\boldsymbol{\mu}$ over functions on $\mathcal{M}$, supported on $L^2(\gamma)$.

2.1.1. Continuum Prior

We assume a Gaussian *prior* distribution $\boldsymbol{\pi}$ on the unknown initial condition $u \in L^2(\gamma)$:

$$\boldsymbol{\pi} = N(0, \mathcal{C}_u), \qquad \mathcal{C}_u = (\alpha I - \Delta_{\mathcal{M}})^{-s/2}, \tag{6}$$

where $\alpha \geq 0$, $s > m$ and $\Delta_{\mathcal{M}}$ denotes the Laplace Beltrami operator. Equation (6) corresponds to the covariance operator description of the Gaussian measure $\boldsymbol{\pi}$. The covariance function representation may be advantageous in the derivation of regression formulae —see the appendix. As described for instance in Gao et al. (2019), the Laplace Beltrami operator is a natural object to define Gaussian processes on manifolds, because its eigenfunctions contain rich geometric information. To provide further intuition, note that draws $u \sim \boldsymbol{\pi}$ can be obtained via the Karhunen-Loève expansion

$$u(x) = \sum_{i=1}^{\infty} (\alpha + \lambda_i)^{-s/4} \xi_i \psi_i(x), \qquad \xi_i \overset{\text{i.i.d}}{\sim} N(0,1), \tag{7}$$

showing that the prior $\boldsymbol{\pi}$ favors functions that have larger components in the first eigenfunctions of $\Delta_{\mathcal{M}}$. The condition $s > m$ guarantees that the expected $L^2(\gamma)$ norm of $u \sim \boldsymbol{\pi}$, which agrees with $\sum_{i=1}^{\infty} (\alpha + \lambda_i)^{-s/2}$, is finite. This in turn implies that $u \sim \boldsymbol{\pi}$ belongs to $L^2(\gamma)$ almost surely. More generally, the parameter s characterizes the almost sure Hölder and Sobolev regularity of draws from $\boldsymbol{\pi}$ (Dashti and Stuart); larger values of s correspond to smoother prior draws. The parameter α gives an effective prior length-scale: frequencies corresponding to $\lambda_i \ll \alpha$ have substantial contribution in the sum in equation (7).

2.1.2. Continuum Forward and Observation Maps

In what follows we take, for concreteness and motivated by applications in image deblurring, the forward map $\mathcal{F} = \mathcal{F}^t$ to be the solution of the heat equation on $\mathcal{M}$ up to a given time

$t \geq 0$. That is, we set

$$\mathcal{F}u \equiv \mathcal{F}^t u := e^{t\Delta_{\mathcal{M}}} u. \tag{8}$$

Note that $\mathcal{M}$ plays two roles in definition of $\mathcal{F}^t$: it determines both the physical domain of the heat equation and the Laplace Beltrami operator $\Delta_{\mathcal{M}}$. Our choice of forward map $\mathcal{F}^t$ includes the identity map (corresponding to regression) as a particular case (for $t = 0$) and gives us the opportunity to study slightly more general data models. We note that $\mathcal{F}^t$ has a natural graph counterpart (see (16)).

We now describe our choice and interpretation of observation maps. Let $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathcal{M}$, and let $\delta > 0$ be small. For $w \in L^2(\gamma)$ we define the j-th coordinate of the vector $\mathcal{O}w$ by

$$[\mathcal{O}w]_j := \frac{1}{\gamma\big(B_\delta(\mathbf{x}_j) \cap \mathcal{M}\big)} \int_{B_\delta(\mathbf{x}_j)\cap\mathcal{M}} w(x)\gamma(dx), \quad 1 \leq j \leq p, \tag{9}$$

where $B_\delta(\mathbf{x}_j)$ denotes the Euclidean ball of radius δ centered at $\mathbf{x}_j$. At an intuitive level, and in our numerical investigations, we see $\mathcal{O}$ as the point-wise evaluation map at the inputs $\mathbf{x}_j$:

$$\mathcal{O}w = [w(\mathbf{x}_1), \ldots, w(\mathbf{x}_p)]' \in \mathbb{R}^p.$$

Henceforth we denote $\mathcal{G} := \mathcal{O} \circ \mathcal{F}$.

Remark 1 *It would be perhaps more intuitive to work with an observation map defined by pointwise evaluations rather than local averages at a certain length-scale δ. Indeed, typically one assumes that the observations y correspond to noisy versions of "true" labels associated to given feature vectors. However, for technical reasons when going from discrete to continuum in the next sections, in the very low number of observed labels regime that we work on (i.e. p does not grow to infinity with n) definition 9 allows us to perform rigorous analysis in an L^2 sense, while pointwise evaluation does not. It is still an open problem to establish uniform type convergence results for eigenvectors of graph Laplacians towards continuum counterparts in the random geometric graph setting; such technical results would allow us to work with the more standard setting for the observation map.*

Having said this, when the continuum prior $\boldsymbol{\pi}$ is supported on a space of regular functions (as is the case when s in (6) is large enough), the posterior (as defined in 11) converges in the limit $\delta \to 0$ to a posterior obtained from a likelihood where the observation map was based on pointwise evaluations. Thus, for strong priors we do not expect much difference between working with one observation model or the other.

2.1.3. Data and Noise Models

Having specified the forward and observation maps $\mathcal{F}$ and $\mathcal{O}$, we assume that the label vector $y \in \mathbb{R}^p$ arises from noisy measurement of $\mathcal{O} \circ \mathcal{F}(u) \in \mathbb{R}^p$. A noise-model will be specified via a function $\phi^y : \mathbb{R}^p \to \mathbb{R}$. We postpone the precise statement of assumptions on ϕ^y to section 4. Two guiding examples, covered by the theory, are given by

$$\phi^y(w) := \frac{1}{2\sigma^2}|y - w|^2, \quad \text{or} \quad \phi^y(w) := -\sum_{i=1}^{p} \log\Big(\Psi\big(y_i w_i; \sigma\big)\Big), \tag{10}$$

where Ψ denotes the CDF of a centered univariate Gaussian with variance σ^2. The former noise model corresponds to Gaussian i.i.d. noise in the observation of each of the p coordinates of $\mathcal{G}u$. The latter corresponds to probit classification, and a noise model of the form $y_i = S\big(v_i + \eta_i\big)$ with η_i i.i.d. $N(0, \sigma^2)$, and S the sign function. For label inference in Bayesian classification, the posterior obtained below needs to be pushed-forward via the sign function (Bertozzi et al., 2018).

2.1.4. Continuum Posterior

The Bayesian solution to the ground-truth continuum learning problem is a continuum posterior measure

$$\begin{aligned}\boldsymbol{\mu}(du) &\propto \exp\big(-\phi^y(\mathcal{G}u)\big)\boldsymbol{\pi}(du)\\ &=: \exp\big(-\Phi(u;y)\big)\boldsymbol{\pi}(du),\end{aligned}\tag{11}$$

that represents the conditional distribution of u given the data y. Equation (11) defines the negative log-likelihood function Φ, that characterizes the conditional distribution of labels y given u. The posterior $\boldsymbol{\mu}$ contains all the information on the unknown input u available in the prior and the data.

2.2. Discrete Learning Setting

We consider the learning of functions defined on a point cloud $\mathcal{M}_n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathcal{M}$. The underlying manifold $\mathcal{M}$ is assumed to be unknown. We suppose to have access to the same label data y as in the continuous setting, and that the inputs $\mathbf{x}_1, \dots, \mathbf{x}_p$ in the definition of $\mathcal{O}$ correspond to the first p points in $\mathcal{M}_n$. Thus, in a physical analogy the data may be interpreted as noisy measurements of the true temperature at the first p points in the cloud at time $t \geq 0$. The aim is to construct —without knowledge of $\mathcal{M}$— a posterior measure $\boldsymbol{\mu_n}$ over functions in $\mathcal{M}_n$ representing the initial temperatures at each point in the cloud.

Similar to the continuous setting, we will denote by $L^2(\gamma_n)$ the space of functions on the cloud that are square integrable with respect to the uniform measure γ_n on $\mathcal{M}_n$. It will be convenient to view, formally, functions $u_n \in L^2(\gamma_n)$ as vectors in $\mathbb{R}^n$. We then write $u_n \equiv [u_n(1), \dots, u_n(n)]'$, and think of $u_n(i)$ as evaluation of the function u_n at $\mathbf{x}_i$.

The graph-posteriors are built by introducing a graph-based prior, and graph-based forward and observation maps $\mathcal{F}_n : L^2(\gamma_n) \to L^2(\gamma_n)$ and $\mathcal{O}_n : L^2(\gamma_n) \to \mathbb{R}^p$. The same noise-model and data as in the continuum case will be used. We start by introducing a graph structure in the point cloud. Graph-based priors and forward maps are defined via a graph-Laplacian that summarizes the geometric information available in the point cloud $\mathcal{M}_n$.

2.2.1. Geometric Graph and Graph-Laplacian

We endow the point cloud with a graph structure. We focus on ε-neighborhood graphs: an input is connected to every input within a distance of ε. A popular alternative are k-nearest neighbor graphs, where an input is connected to its k nearest neighbors. The influence of the choice of graphs in unsupervised learning is studied in Maier et al. (2009).

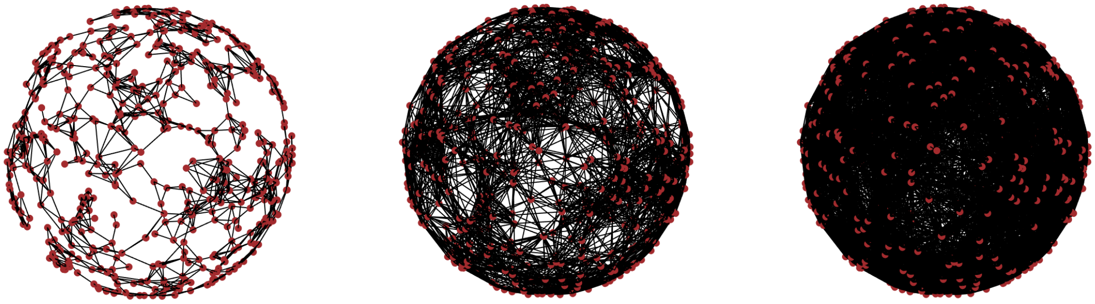

Figure 1: Geometric graphs with $n = 500$, and $\varepsilon = n^{-1/4}, 2n^{-1/4}$, and $3n^{-1/4}$ from left to right.

First, consider the kernel function $K : [0, \infty) \to [0, \infty)$ defined by

$$K(r) := \begin{cases} 1 & \text{if } r \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

For $\varepsilon > 0$ we let $K_\varepsilon : [0, \infty) \to [0, \infty)$ be the rescaled version of K given by

$$K_\varepsilon(r) := \frac{m+2}{n^2 \alpha_m \varepsilon^{m+2}} K\left(\frac{r}{\varepsilon}\right),$$

where α_m denotes the volume of the m-dimensional unit ball. We then define the weight $W_n(\mathbf{x}_i, \mathbf{x}_j)$ between $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{M}_n$ by

$$W_n(\mathbf{x}_i, \mathbf{x}_j) := K_{\varepsilon_n}(|\mathbf{x}_i - \mathbf{x}_j|),$$

for a given choice of parameter $\varepsilon = \varepsilon_n$, where we have made the dependence of the *connectivity rate* ε on n explicit. In order for the graph-based learning problems to be consistent in the large n limit, ε should be scaled appropriately with n —see subsection 4.1. Figure 1 shows three geometric graphs $(\mathcal{M}_n, W_n)$ with fixed n and different choices of connectivity ε.

We now define the *graph Laplacian* of the geometric graph $(\mathcal{M}_n, W_n)$ by

$$\Delta_{\mathcal{M}_n} := D_n - W_n, \tag{13}$$

where D is the degree matrix of the weighted graph, i.e., the diagonal matrix with diagonal entries $D_{ii} = \sum_{j=1}^n W_n(\mathbf{x}_i, \mathbf{x}_j)$. Several definitions of graph Laplacian co-exist in the literature; the one above is some times referred to as the *unnormalized* graph Laplacian Von Luxburg (2007). As will be made precise, the performance of the learning methods considered here is largely determined by the behavior of the spectrum of the graph Laplacian. Throughout we denote its eigenpairs by $\{\lambda_i^n, \psi_i^n\}_{i=1}^n$, and assume that the eigenvalues are in non-decreasing order.

2.2.2. Graph Prior

A straight-forward discrete analogue to (6) suggests endowing the unknown function u_n with a prior

$$\widetilde{\boldsymbol{\pi_n}} = N(0, \mathcal{C}_{u_n}), \qquad \mathcal{C}_{u_n} := (\alpha I_n + \Delta_{\mathcal{M}_n})^{-s/2}, \tag{14}$$

where $\alpha \geq 0$ and $s > m$ are chosen as in (6). Like the continuum prior, the graph-based one favors functions u_n with large components in the first eigenfunctions of $\Delta_{\mathcal{M}_n}$, thus infusing geometric information on the probabilistic Bayesian reconstruction (Bertozzi et al., 2018). The graph Laplacian, in contrast to the regular Laplacian, is positive semi-definite, and hence the change in sign with respect to (6). This choice of graph prior was considered in García Trillos and Sanz-Alonso (2018a), and also in Bertozzi et al. (2018) in the case $\alpha = 0, s = 2$. In this paper we introduce and study priors $\boldsymbol{\pi_n}$ defined in terms of truncation of the priors $\widetilde{\boldsymbol{\pi_n}}$, retaining only the portion of the spectra of $\Delta_{\mathcal{M}_n}$ that provably approximates that of $-\Delta_{\mathcal{M}}$.

Precisely, we define the graph priors $\boldsymbol{\pi_n}$ as the law of u_n given by

$$u_n = \sum_{i=1}^{k_n} (\alpha + \lambda_i^n)^{-s/4} \xi_i \psi_i^n, \qquad \xi_i \overset{\text{i.i.d}}{\sim} N(0,1), \tag{15}$$

where $k_n \leq n$ may be chosen freely with the restrictions that $k_n \to \infty$ and $\lim_{n\to\infty} k_n \varepsilon_n^m = 0$. Such choice is possible as long as the connectivity ε_n decays with n.

2.2.3. Graph Forward and Observation Maps

We define a forward map $\mathcal{F}_n : L^2(\gamma_n) \to L^2(\gamma_n)$ by

$$\mathcal{F}_n u_n \equiv \mathcal{F}_n^t u_n := e^{-t\Delta_{\mathcal{M}_n}} u_n, \tag{16}$$

where $t \geq 0$ is given as in the continuum setting. Likewise, for $\delta > 0$ as in (9) we define an observation map $\mathcal{O}_n : L^2(\gamma_n) \to \mathbb{R}^p$ by

$$[\mathcal{O}_n w](j) := \frac{1}{n\gamma_n\big(B_\delta(\mathbf{x}_j)\big)} \sum_{k:\, \mathbf{x}_k \in B_\delta(\mathbf{x}_j) \cap \mathcal{M}_n} w(k), \quad 1 \leq j \leq p.$$

As in the continuum setting, $\mathcal{O}_n$ should be thought of as point-wise evaluation at the inputs $\{\mathbf{x}_i\}_{i=1}^p$ and we denote $\mathcal{G}_n := \mathcal{O}_n \circ \mathcal{F}_n$.

2.2.4. Data and Likelihood

For the construction of graph posteriors we use the same labeled data y and noise model $\phi^y : \mathbb{R}^p \to \mathbb{R}$ as in the continuum case —see subsection 2.1.3.

2.2.5. Graph Posterior

We define the *graph-posterior* measure $\boldsymbol{\mu_n}$ by

$$\begin{aligned} \boldsymbol{\mu_n}(du) &\propto \exp\big(-\phi^y(\mathcal{G}_n u_n)\big)\boldsymbol{\pi_n}(du_n) \\ &=: \exp\big(-\Phi_n(u_n; y)\big)\boldsymbol{\pi_n}(du_n), \end{aligned} \tag{17}$$

where $\boldsymbol{\pi_n}$ is the (truncated) graph prior defined as the law of (15), and the above expression defines the function Φ_n, interpreted as a graph-based approximation to the negative log-likelihood.

In subsection 4.1 we will contrast the above “truncated" graph-posteriors to the “untruncated" graph-posteriors

$$\begin{aligned} \widetilde{\boldsymbol{\mu_n}}(du) &\propto \exp\bigl(-\phi^y(\mathcal{G}_n u_n)\bigr)\widetilde{\boldsymbol{\pi_n}}(du_n) \\ &=: \exp\bigl(-\Phi_n(u_n;y)\bigr)\widetilde{\boldsymbol{\pi_n}}(du), \end{aligned} \tag{18}$$

obtained by using the prior $\widetilde{\boldsymbol{\pi_n}}$ in equation (14).

3. Posterior Sampling: pCN and Graph-pCN

The continuum limit theory developed in García Trillos and Sanz-Alonso (2018a) and recalled in subsection 4.1 suggests viewing graph posteriors $\boldsymbol{\mu_n}$ as discretizations of a posterior measure over functions on the underlying manifold. Again, these discretizations are robust for fixed p and growing number of total inputs n. This observation substantiates the idea introduced in Bertozzi et al. (2018) of using a version of the pCN MCMC method (Beskos et al., 2008) for robust sampling of graph posteriors. We review the continuum pCN method in subsection 3.1, and the graph pCN counterpart in subsection 3.2.

3.1. Continuum pCN

In practice, sampling of functions on the continuum always requires a discretization of the infinite dimensional function, usually defined in terms of a mesh and possibly a series truncation. A fundamental idea is that algorithmic robustness with respect to discretization refinement can be guaranteed by ensuring that the algorithm is well defined in function space, before discretization (Dashti and Stuart). This insight led to the formulation of the pCN method for sampling of conditioned diffusions (Beskos et al., 2008), and of measures arising in Bayesian nonparametrics in Cotter et al. (2009). The pCN method for sampling the continuum posterior measure (11) is given in Algorithm 1.

Algorithm 1 Continuum pCN

```
Set $j = 0$ and pick any $u^{(0)} \in L^2(\gamma)$.
Propose $\tilde{u}^{(j)} = (1-\beta^2)^{1/2} u^{(j)} + \beta \zeta^{(j)}$,    where $\zeta^{(j)} \sim N(0, \mathcal{C}_u)$.
Set $u^{(j+1)} = \tilde{u}^{(j)}$ with probability
    $a\bigl(u^{(j)}, \tilde{u}^{(j)}\bigr) := \min\Bigl\{1, \exp\Bigl(\Phi\bigl(u^{(j)}; y\bigr) - \Phi\bigl(\tilde{u}^{(j)}; y\bigr)\Bigr)\Bigr\}.$
Set $u^{(j+1)} = u^{(j)}$ otherwise.
$j \to j+1$.
```

Posterior expectations of suitable test functions f can then be approximated by empirical averages

$$\boldsymbol{\mu}(f) \approx \frac{1}{J}\sum_{j=1}^{J} f\big(u^{(j)}\big) = S^J(f). \tag{19}$$

The user-chosen parameter $\beta \in [0,1]$ in Algorithm 1 monitors the step-size of the chain jumps: larger β leads to larger jumps, and hence to more state space exploration, more rejections, and slower probing of high probability regions. Several robust discretization properties of Algorithm 1 —that contrast with the deterioration of traditional random walk approaches— have been proved in Hairer et al. (2014). Note that the acceptance probability is determined by the potential Φ (here interpreted as the negative log-likelihood) that defines the density of the posterior with respect to the prior. In the extreme case where Φ is constant, moves are always accepted. However, if the continuum posterior is far from the continuum prior, the density will be far from constant. This situation may arise, for instance, in cases where p is large or the size σ of the observation noise is small. A way to make posterior informed proposals that may lead to improved performance in these scenarios has been proposed in Rudolf and Sprungk (2015).

3.2. Graph pCN

The graph pCN method is described in Algorithm 2, and is defined in complete analogy to the continuum pCN, Algorithm 1. When considering a sequence of problems with fixed p and increasing n, the continuum theory intuitively supports the robustness of the method. Moreover, as indicated in Bertozzi et al. (2018) the parameter β may be chosen independently of the value of n. Our experiments in section 5 confirm this robustness, and also investigate the deterioration of the acceptance rate when both n and p are large.

Algorithm 2 Graph pCN

```
  Set $j = 0$ and pick any $u_n^{(0)} \in L^2(\gamma_n)$.
  Propose $\tilde{u}_n^{(j)} = (1-\beta^2)^{1/2} u_n^{(j)} + \beta \zeta_n^{(j)}$,    where $\zeta_n^{(j)} \sim N(0, \mathcal{C}_{u_n})$.
  Set $u_n^{(j+1)} = \tilde{u}_n^{(j)}$ with probability

        $a_n\big(u_n^{(j)}, \tilde{u}_n^{(j)}\big) := \min\Big\{1, \exp\Big(\Phi_n\big(u_n^{(j)}; y\big) - \Phi_n\big(\tilde{u}_n^{(j)}; y\big)\Big)\Big\}.$

  Set $u_n^{(j+1)} = u_n^{(j)}$ otherwise.
  $j \to j+1$.
```

Again, graph-posterior expectations of suitable test functions f_n can then be approximated by empirical averages

$$\widetilde{\boldsymbol{\mu}_{\boldsymbol{n}}}(f_n) \approx \frac{1}{J}\sum_{j=1}^{J} f_n\big(u_n^{(j)}\big) = S^J(f_n). \tag{20}$$

In *informal* but intuitive terms, the uniform spectral gap that we establish below shows that the large J asymptotic variance of $S^J(f_n)$ is independent of n.

4. Main Results

4.1. Continuum Limits

The paper García Trillos and Sanz-Alonso (2018a) established large n asymptotic convergence of the untruncated graph-posteriors $\widetilde{\boldsymbol{\mu}_n}$ in (18) to the continuum posterior $\boldsymbol{\mu}$ in (11). The convergence was established in a topology that combines Wasserstein distance and an L^2-type term in order to compare measures over functions in the continuum with measures over functions in graphs.

Proposition 2 (Theorem 4.4 in García Trillos and Sanz-Alonso (2018a)) *Suppose that $s > 2m$ and that*

$$\frac{(\log(n))^{p_m}}{n^{1/m}} \ll \varepsilon_n \ll \frac{1}{n^{1/s}}, \qquad \text{as } n \to \infty, \tag{21}$$

where $p_m = 3/4$ for $m = 2$ and $p_m = 1/m$ for $m \geq 3$. Then, the untruncated graph-posteriors $\widetilde{\boldsymbol{\mu}_n}$ converge towards the posterior $\boldsymbol{\mu}$ in the $\mathcal{P}(TL^2)$ sense.

We refer to Appendix B for the construction of the metric space TL^2 that was originally introduced in García Trillos and Slepčev (2016a). Notice that in the space TL^2 we can compare functions defined on $\mathcal{M}_n$ with functions defined on $\mathcal{M}$. The space $\mathcal{P}(TL^2)$ was introduced in García Trillos and Sanz-Alonso (2018a) and stands for the set of Borel probability measures on TL^2 endowed with the topology of weak convergence. This space allows us to formalize the convergence of a sequence of probability distributions over functions on $\mathcal{M}_n$ to a probability distribution over functions on $\mathcal{M}$. In particular, in the previous theorem, convergence is interpreted as: $\widetilde{\boldsymbol{\mu}_n}$ converges weakly to $\boldsymbol{\mu}$ as $n \to \infty$, all measures seen as elements of $\mathcal{P}(TL^2)$. It is important to note that in the theorem, convergence refers to the limit of *fixed* labeled data set of size p, and growing size of unlabeled data. In order for the continuum limit to hold, the connectivity of the graph ε_n needs to be carefully scaled with n as in (21).

At an intuitive level, the lower bound on ε_n guarantees that there is enough averaging in the limit to recover a meaningful deterministic quantity. The upper bound ensures that the graph priors converge appropriately towards the continuum prior. At a deeper level, the lower bound is an order one asymptotic estimate for the ∞-optimal transport distance between the uniform and uniform empirical measure on the manifold (García Trillos and Slepčev, 2014), that hinges on the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ lying *on* the manifold $\mathcal{M}$: if the inputs were sampled from a distribution whose support is close to $\mathcal{M}$, but whose intrinsic dimension is d and not m, then the lower bound would be written in terms of d instead of m. The upper bound, on the other hand, relies on the approximation bounds (24) of the continuum spectrum of the Laplace-Beltrami by the graph Laplacian.

We now present a new result on the stability of intrinsically constructed posteriors, generalized to $\mathcal{M}$ by interpolation via the map $\mathcal{I}_n := \mathcal{I}_n^1$ —see (4); this is the most basic interpolation map that can be constructed exclusively from the point cloud $\mathcal{M}_n$ and the metric on the ambient space. Other than extending the theory to cover the important question of generalization, there is another layer of novelty in Theorem 3: graph-posteriors are constructed with truncated priors, and the upper-bound in the connectivity ε_n in (21) is removed. As discussed in subsection 5.1.1, only a portion of the spectrum of the graph Laplacian contains relevant information of the underlying manifold $\mathcal{M}$, and thus nothing is lost by throwing away higher modes. See Figure 2 for an illustration.

Theorem 3 *Suppose that $s > 2m$ and that*

$$\frac{(\log(n))^{p_m}}{n^{1/m}} \ll \varepsilon_n \ll 1, \qquad \text{as } n \to \infty, \tag{22}$$

where p_m is as in Proposition 2. Then, with probability one,

$$\mathcal{I}_{n\sharp}\boldsymbol{\mu_n} \to_{\mathcal{P}(L^2(\gamma))} \boldsymbol{\mu}, \qquad \text{as } n \to \infty.$$

The proof is presented in Appendix C. Similar results hold for more general interpolation maps as long as they are uniformly controlled and consistent when evaluated at the eigenfunctions of graph Laplacians (see Remark 13).

Remark 4 *Our results concern the regime where $n \to \infty$ and p is constant. This corresponds to the semi-supervised setting of many more unlabeled data points than labels. Our analysis would also allow us to take the double limit $n \to \infty$ followed by $p \to \infty$. This corresponds to a semi-supervised learning regime where both the number of unlabeled data points and the number of labeled data points grow, but p grows at the slowest rate possible. In that regime the limiting posterior concentrates around a single function on $\mathcal{M}$ which would correspond to the true “regression function". It may be possible to establish similar posterior concentration results in the regime where both $n \to \infty$ and $p = p_n \to \infty$ go simultaneously to infinity as well as to establish posterior contraction rates. We leave such analysis for future work.*

4.2. Uniform Spectral Gaps for Graph-pCN Algorithms

The aim of this subsection is to establish how, in a precise and rigorous sense, the graph-pCN method in Algorithm 2 is insensitive to the increase of the number n of input data provided that the number p of labeled data is fixed and that a continuum limit exists. This behavior contrasts dramatically with other sampling methodologies such as the random walk sampler. One could characterize the robustness of MCMC algorithms in terms of uniform spectral gaps.

We start by defining the spectral gap for a single Markov chain with state space an arbitrary separable Hilbert space $\mathcal{H}$. We consider two notions of spectral gap, one using Wasserstein distance with respect to some distance like function $\tilde{d}$, and the other one in terms of L^2. For the purposes of this paper the Wasserstein spectral gap can be thought as an intermediate step which is “easier" to prove directly following the ideas introduced in Hairer et al. (2014), while the L^2 gap is a consequence whose implications are meaningful for our problem. We start with the two definitions.

Definition 5 (Wasserstein spectral gaps) *Let P be the transition kernel for a discrete time Markov chain with state space $\mathcal{H}$. Let $\tilde{d} : \mathcal{H} \times \mathcal{H} \to [0, 1]$ be a distance like function* *(i.e. a symmetric, lower-semicontinuous function satisfying $\tilde{d}(u,v)=0$ if and only if $u=v$). Without the loss of generality we also denote by $\tilde{d}$ the Wasserstein distance (1-OT distance) on $\mathcal{P}(\mathcal{H})$ induced by $\tilde{d}$ (see (34)). We say that P has spectral gap if there exist positive constants C,λ such that*

$$\tilde{d}(P^j\mu, P^j\nu) \leq C\exp(-\lambda j)\tilde{d}(\mu,\nu), \quad \forall \mu,\nu \in \mathcal{P}(\mathcal{H}), \quad \forall j \in \mathbb{N}.$$

In the above $\mathcal{P}(\mathcal{H})$ stands for the set of Borel probability measures on $\mathcal{H}$.

Definition 6 (L^2-spectral gaps) *Let P be the transition kernel for a discrete time Markov chain with state space $\mathcal{H}$ and suppose that μ is invariant under P. P is said to have L^2_μ-spectral gap $1-\exp(-\lambda)$ (for $\lambda>0$) if for every $f \in L^2(\mathcal{H};\mu)$ we have*

$$\frac{\|Pf-\mu(f)\|^2_{L^2(\mathcal{H};\mu)}}{\|f-\mu(f)\|^2_{L^2(\mathcal{H};\mu)}} \leq \exp(-\lambda).$$

In the above, $\mu(f) := \int_{\mathcal{H}} f(u)d\mu(u)$ and $Pf(u) := \int_{\mathcal{H}} f(v)P(u,dv)$.

Having defined the notion of spectral gap for a single Markov chain, the notion of uniform spectral gap for a family of Markov chains is defined in an obvious way. Namely, if $\{P_n\}_{n\in\mathbb{N}}$ is a family of Markov chains, with perhaps different state spaces $\{\mathcal{H}_n\}_{n\in\mathbb{N}}$, we say that the family of Markov chains has *uniform* Wasserstein spectral gap with respect to a family of distance like functions $\{\tilde{d}_n\}$ if the Markov chains have spectral gaps with constants C,λ which can be uniformly bounded, respectively, from above and away from zero. Likewise the chains are said to have uniform L^2-gaps (with respect to respective invariant measures) if the constant λ can be uniformly bounded away from zero. We remark that Wasserstein spectral gaps imply uniqueness of invariant measures of Markov chains (this follows directly from the definition of Wasserstein gap).

Having introduced the above notions of “mixing" for Markov chains in a general setting, we return to the problem of understanding the mixing of the family of pCN algorithms for our semi-supervised learning problem. We will make the following assumption on the negative log-likelihood function ϕ^y.

Assumption 1 *Let $\beta \in (0,1]$. For a certain fixed $y \in \mathbb{R}^p$ we assume the following conditions on $\phi^y : \mathbb{R}^p \to \mathbb{R}$.*

i) *For every $K>0$ there exists $c \in \mathbb{R}$ such that if $v,w \in \mathbb{R}^p$ satisfy*

$$|w-\sqrt{1-\beta^2}\, v| \leq K$$

then,

$$\phi^y(v)-\phi^y(w) \geq c.$$

ii) *(Linear growth of local Lipschitz constant) There exists a constant L such that*

$$|\phi^y(v)-\phi^y(w)| \leq L\max\{|v|,|w|,1\}|v-w|, \quad \forall v,w \in \mathbb{R}^p.$$

In Appendix E we show that the Gaussian model and the probit model satisfy these assumptions.

In what follows it is convenient to use $\mathcal{H}$ as a placeholder for one of the spaces $L^2(\gamma_n)$, $n \in \mathbb{N}$, or the space $L^2(\gamma)$. Likewise P is a placeholder for the transition kernel associated to the pCN scheme from section 3 defined on $\mathcal{H}$ for each choice of $\mathcal{H}$. We are ready to state our second main theorem:

Theorem 7 (Uniform Wasserstein spectral gap) *Let $\theta > 0$, $\eta > 0$. For each choice of $\mathcal{H}$ let $d : \mathcal{H} \times \mathcal{H} \to [0, 1]$,*

$$d(u, v) := \min\Big\{1, \frac{\overline{d}(u, v)}{\theta}\Big\}, \quad u, v \in \mathcal{H}$$

be a rescaled and truncated version of the distance

$$\overline{d}(u, v) := \inf_{T, \psi \in A(T,u,v)} \int_0^T \exp(\eta\|\psi\|)dt,$$

$$A(T, u, v) := \{\psi \in C^1([0, T]; \mathcal{H}) \ : \ \psi(0) = u, \quad \psi(T) = v, \quad \|\dot{\psi}\| = 1\}.$$

Finally, let $\tilde{d}$ be the distance-like function

$$\tilde{d}(x, y) := \sqrt{d(x, y)(1 + V(x) + V(y))}, \quad u, v \in \mathcal{H}$$

where

$$V(u) := \|u\|^2, \quad u \in \mathcal{H}.$$

Then, under the assumptions of Theorem 3 and Assumption 1, $\theta > 0$ and $\eta > 0$ can be chosen independently of the specific choice of $\mathcal{H}$ in such a way that

$$\tilde{d}(P^j\nu_1, P^j\nu_2) \leq C \exp(-\lambda j)\tilde{d}(\nu_1, \nu_2), \quad \forall \nu_1, \nu_2 \in \mathcal{P}(\mathcal{H}), \quad \forall j \in \mathbb{N},$$

for constants C, λ that are independent of the choice of $\mathcal{H}$.

A few remarks help clarify our results.

Remark 8 *Notice that $\overline{d}$ is a Riemannian distance whose metric tensor changes in space and takes larger values for points that are far away from the origin (notice that the choice $\eta = 0$ returns the canonical distance on $\mathcal{H}$). In particular, points that are far away from the origin have to be very close in the canonical distance in order to be close in the d distance. This distance was considered in Hairer et al. (2014). We would also like to point out that the exponential form of the metric tensor can be changed to one with polynomial growth given the choice of V.*

Remark 9 *Theorem 7 is closely related to Theorem 2.14 in Hairer et al. (2014). There, uniform spectral gaps are obtained for the family of pCN kernels indexed by the truncation levels of the Karhunen Loève expansion of the continuum prior. For that type of discretization, all distributions are part of the same space; this contrasts with our set-up where the discretizations of the continuum prior are the graph priors.*

Due to the reversibility of the kernels associated to the pCN algorithms (they are particular instances of Metropolis-Hastings), Theorem 7 implies uniform L^2-spectral gaps as introduced earlier. Notice that the Wasserstein gaps imply uniqueness of invariant measures (which are precisely the graph and continuum posteriors for each setting) and hence there is no ambiguity when talking about L^2-spectral gaps.

Corollary 10 *Under the assumptions of Theorem 3 and Assumption 1 the kernel associated to the pCN algorithm has an L^2-spectral gap independent of the choice of $\mathcal{H}$.*

The proof of Theorem 7 and its corollary are presented in Appendix D.

Recall that graph-posterior expectations of suitable test functions f_n can be approximated by empirical averages

$$\widetilde{\boldsymbol{\mu}_{\boldsymbol{n}}}(f_n) \approx \frac{1}{J}\sum_{j=1}^{J} f_n\big(u_n^{(j)}\big) = S^J(f_n). \tag{23}$$

Roughly speaking, this uniform spectral gap shows that the large J asymptotic variance of $S^J(f_n)$ is independent of n. Uniform spectral gaps may be used to find uniform bounds on the asymptotic variance of empirical averages (Kipnis and Varadhan, 1986).

Remark 11 *It is important to highlight that the uniform gaps for the pCN algorithm (when n grows) depend nonetheless on the number of observations p, and that the gaps may collapse with growing p. This should be intuitively reasonable as this corresponds to considering a more complex likelihood function, which in turn pushes the posterior further from the prior.*

5. Numerical Study

In the numerical experiments that follow we take $\mathcal{M} = \mathcal{S}$ to be the two-dimensional sphere in $\mathbb{R}^3$. Our main motivation for this choice of manifold is that it allows us to expediently make use of well-known closed formulae (Olver, 2013) for the spectrum of the spherical Laplacian $\Delta_{\mathcal{M}} = \Delta_{\mathcal{S}}$ in the continuum setting that serves as our ground truth model. We recall that $-\Delta_{\mathcal{S}}$ admits eigenvalues $l(l+1)$, $l \geq 0$, with corresponding eigenspaces of dimension $2l+1$. These eigenspaces are spanned by spherical harmonics (Olver, 2013). In subsections 5.1, 5.2, and 5.3 we study, respectively, the spectrum of graph Laplacians, continuum limits, and the scalability of pCN methods.

5.1. Spectrum of Graph Laplacians

The asymptotic behavior of the spectra of graph-Laplacians is crucial in the theoretical study of consistency of graph-based methods. In subsection 5.1.1 we review approximation bounds that motivate our truncation of graph-priors, and in subsection 5.1.2 we comment on the theory of regularity of functions on graphs.

5.1.1. Approximation Bounds

Quantitative error bounds for the difference of the spectrum of the graph Laplacian and the spectrum of the Laplace-Beltrami operator are given in Burago et al. (2014) and García Trillos et al. (2018). Those results imply that, with very high probability,

$$\Big|1 - \frac{\lambda_i^n}{\lambda_i}\Big| \le C\Big(\frac{\delta_n}{\varepsilon_n} + \varepsilon_n\sqrt{\lambda_i}\Big), \quad \forall i, \tag{24}$$

where δ_n denotes the ∞-optimal transport distance (García Trillos and Slepčev, 2014) between the uniform and the uniform empirical measure on the underlying manifold. The important observation here is that the above estimates are only relevant for the first portion of the spectra (in particular for those indices i for which $\varepsilon_n\sqrt{\lambda_i}$ is small). The truncation point at which the estimates stop being meaningful can then be estimated combining (24) and Weyl's formula for the growth of eigenvalues of the Laplace Beltrami operator on a compact Riemannian manifold of dimension m (García Trillos and Sanz-Alonso, 2018a). Namely, from $\lambda_i \sim i^{2/m}$ we see that $\varepsilon_n\sqrt{\lambda_i} \ll 1$ as long as $i = 1, \dots, k_n$ and

$$1 \ll k_n \ll \frac{1}{\varepsilon_n^m}.$$

This motivates our truncation point for graph priors in equation (15).

Figure 2 illustrates the approximation bounds (24). The figure shows the eigenvalues of the graph Laplacian for three different choices of connectivity length scale ε and three different choices of number n of inputs in the graph; superimposed is the spectra of the spherical Laplacian. We notice the flattening of the spectra of the graph Laplacian and, in particular, how the eigenvalues of the graph Laplacian start deviating substantially from those of the Laplace-Beltrami operator after some point in the x-axis. As discussed in García Trillos et al. (2018), the estimates (24) are not necessarily sharp, and may be conservative in suggesting where the deviations start.

5.1.2. Regularity of Discrete Functions

We numerically investigate the role of the parameter s in the discrete regularity of functions $u_n \in L^2(\gamma_n)$ sampled from $\boldsymbol{\pi_n}$. We focus on studying the oscillations of a function within balls of radius ε_n. More precisely, we consider

$$[osc_{\varepsilon_n}(u_n)](\mathbf{x}_i) := \max_{x,z \in B_{\varepsilon_n}(\mathbf{x}_i)\cap \mathcal{M}_n} |u_n(x) - u_n(z)|, \quad i = 1, \dots, n.$$

For given $s = 2, 3, \dots, 8$ we take 100 samples $u_n \sim \boldsymbol{\pi_n}$, and we normalize so that

$$\langle \Delta_n^s u_n, u_n\rangle_{L^2(\gamma_n)} = 1.$$

We then compute the maximum value of $[osc_{\varepsilon_n}(u_n)](\mathbf{x}_i)$ over all $i = 1, \dots, n$ and over all samples u_n and plot the outcome against s. The results are shown in Figure 3. This experiment illustrates the regularity of functions with bounded H_n^s semi-norm

$$\|u_n\|^2_{H_n^s} := \sum_{i=1}^{k_n} (\lambda_i^n)^s \langle u_n, \psi_i^n\rangle^2_{L^2(\gamma_n)}.$$

As expected, higher values of s enforce more regularity on the functions. Notice that here we only consider functions u_n in the support of $\boldsymbol{\pi_n}$ and hence we remove the effect of high

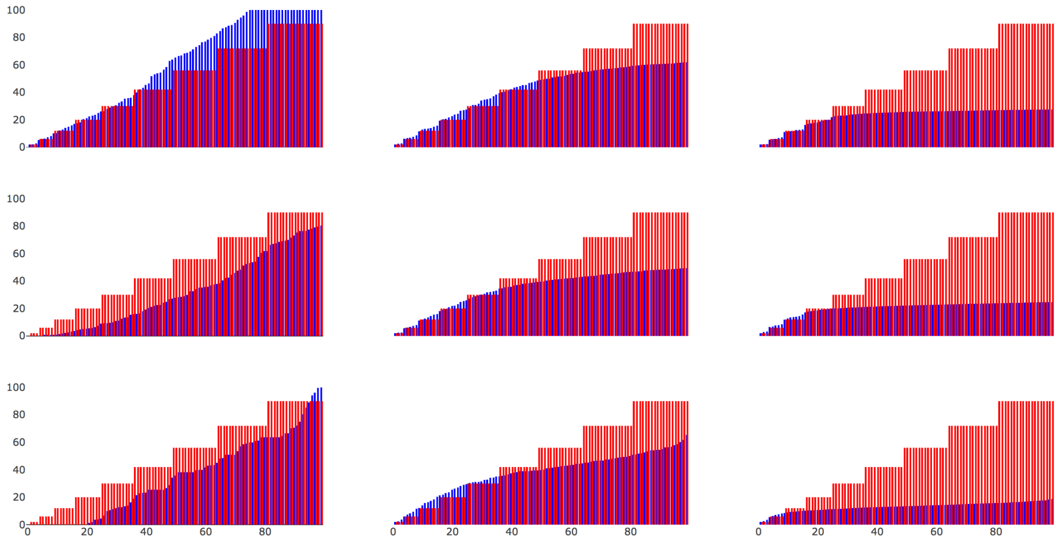

Figure 2: Spectra of spherical and graph Laplacians in red and blue, respectively. Charts are arranged such that ε varies as $[1,\ 2,\ 3] \times n^{-1/4}$ horizontally and n varies as $[1000,\ 500,\ 100]'$ vertically.

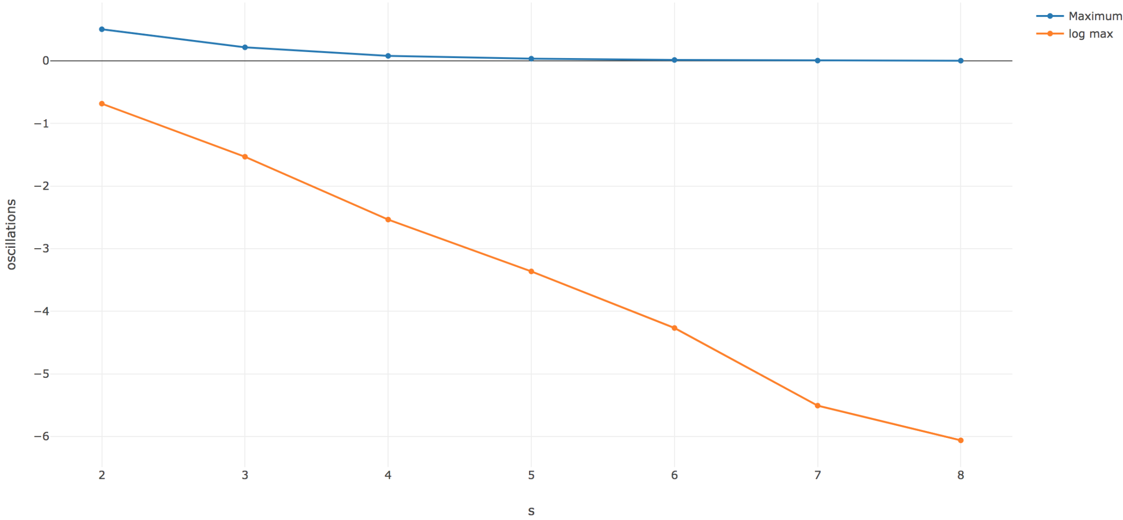

Figure 3: The figure shows the maximum (and its logarithm) amplitude of oscillations for different values of the regularity parameter s.

eigenfunctions of Δ_n (which may be irregular). In particular, the regularity of the functions u_n must come from the regularity of the first eigenvectors of Δ_n together with the growth of $(\lambda_i^n)^s$. To the best of our knowledge nothing is known about regularity of eigenfunctions of graph Laplacians. Studying such regularity properties is an important direction to explore in the future as we believe it would allow us to go beyond the L^2 set-up that we consider for the theoretical results in this paper. In that respect we would like to emphasize that the observation maps considered for the theory of this work are defined in terms of averages and not in terms of pointwise evaluations, but that for our numerical experiments we have used the latter.

A closely related setting in which discrete regularity has been mathematically studied is in the context of *graph* $\mathtt{p}$*-Laplacian semi-norm* (here $\mathtt{p}$ denotes an arbitrary number greater than one, and is not to be confused with the number p of labeled data points). Lemma 4.1 in Slepčev and Thorpe (2017) states that, under the assumptions on ε_n from Theorem 3, for all large enough n and for every discrete function u_n satisfying

$$\mathcal{E}_n^{(\mathtt{p})}(u_n) := \frac{1}{n^2 \varepsilon_n^{\mathtt{p}}} \sum_{i,j} K\left(\frac{|\mathbf{x}_i - \mathbf{x}_j|}{\varepsilon_n}\right) |u_n(\mathbf{x}_i) - u_n(\mathbf{x}_j)|^{\mathtt{p}} = 1,$$

it holds

$$[osc_{\varepsilon_n}(u_n)](\mathbf{x}_i) \leq C^{1/\mathtt{p}}\, n^{1/\mathtt{p}}\, \varepsilon_n, \quad \forall i = 1, \dots, n.$$

This estimate allows to establish uniform convergence (and not simply convergence in TL^2) of discrete functions towards functions defined at the continuum level. More precisely, suppose that $\mathtt{p} > m$ and that $\varepsilon_n \ll \frac{1}{n^{1/\mathtt{p}}}$. Let $\{u_n\}_{n\in\mathbb{N}}$ be a sequence with $u_n \in L^2(\gamma_n)$ converging to a function $u \in L^2(\gamma)$ in the TL^2 sense and for which

$$\sup_{n\in\mathbb{N}} \mathcal{E}_n^{(\mathtt{p})}(u_n) < \infty.$$

Then, u must be continuous (in fact Hölder continuous with Hölder constant obtained from the Sobolev embedding theorem) and moreover

$$\max_{i=1,\dots,n} |u_n(\mathbf{x}_i) - u(\mathbf{x}_i)| \to 0, \quad \text{as } n \to \infty.$$

This is the content of Lemma 4.5 in Slepčev and Thorpe (2017). This type of result rigorously justifies pointwise evaluation of discrete functions with bounded graph $\mathtt{p}$-Laplacian seminorm and the stability of this operation as $n \to \infty$.

5.2. Continuum Limits

5.2.1. Set-up

For the remainder of section 5 we work under the assumption of Gaussian observation noise, so that

$$\Phi(u; y) = \frac{1}{2\sigma^2}|y - \mathcal{G}(u)|^2, \quad \Phi_n(u_n, y) = \frac{1}{2\sigma^2}|y - \mathcal{G}_n(u_n)|^2. \tag{25}$$

The synthetic data y in our numerical experiments is generated by drawing a sample $\eta \sim N(0, \sigma^2 I_{p\times p})$, and setting

$$y = \mathcal{G}(u^\dagger) + \eta,$$

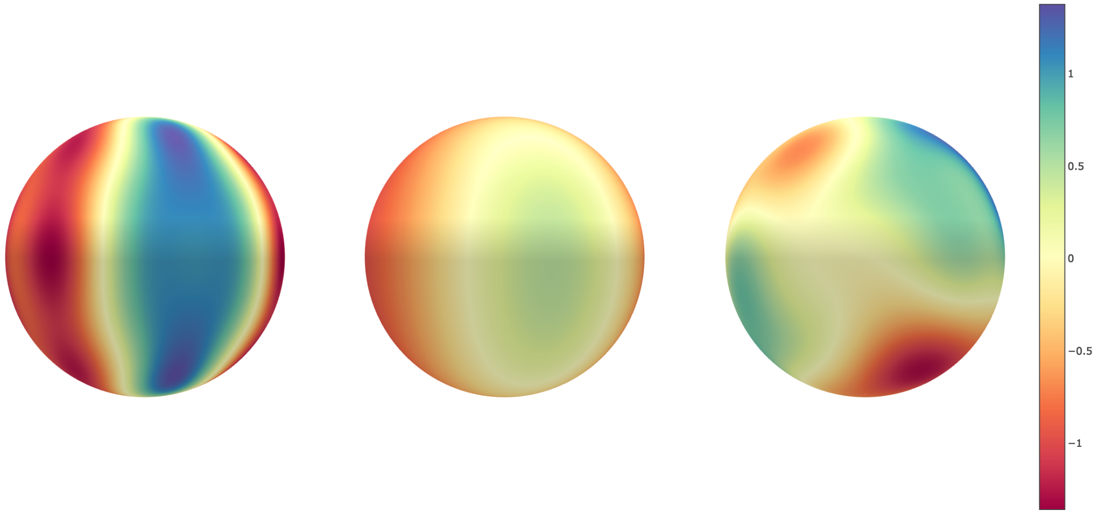

Figure 4: From left to right: Initial condition $u^\dagger$ used as ground truth to generate our synthetic data; heat at $t = 0.5$ with initial condition $u^\dagger$; and draw from the continuous prior.

where $u^\dagger$ is the function in the left panel of Figure 4. We consider several choices of $t \geq 0$, number p of labeled data points, and size of observation noise $\sigma > 0$. The parameters s and α in the prior measures are fixed to $s = 5$, $\alpha = 1$ throughout.

The use of Gaussian observation noise, combined with the linearity of our forward and observation maps, allows us to derive closed formulae for the graph and continuum posteriors. We do so in the the appendix.

5.2.2. Numerical Results

Here we complement the theory by studying the effect that various model parameters have in the accurate approximation of continuum posteriors by graph posteriors. We emphasize that the continuum posteriors serve as a gold standard for our learning problem: graph posteriors built with appropriate choices of connectivity ε result in good approximations to continuum posteriors; however, reconstruction of the unknown function $u^\dagger$ is *not* accurate if the data is not informative enough. In such case, MAPs constructed with graph or continuum posteriors may be far from $u^\dagger$.

All graph-posterior means in the figures are represented using a k-NN interpolation map, as defined in equation (4), with $k = 4$. The posterior means, discrete and continuum, have been obtained using the appropriate pCN algorithm. The pCN algorithm was run for 10^5 iterations, and the last 10^4 samples were used to compute quantities of interest (e.g means and variances). Figure 5 shows a graph-prior draw represented in the point cloud (left), and the associated 4-NN interpolant (right).

Figure 6 shows graph and continuum posteriors with $t = 0$, $t = 0.1$, and $t = 0.3$. For these plots, a suitable choice of graph connectivity ε was taken. In all three cases we see remarkable similarity between the graph and continuum posterior means. However,

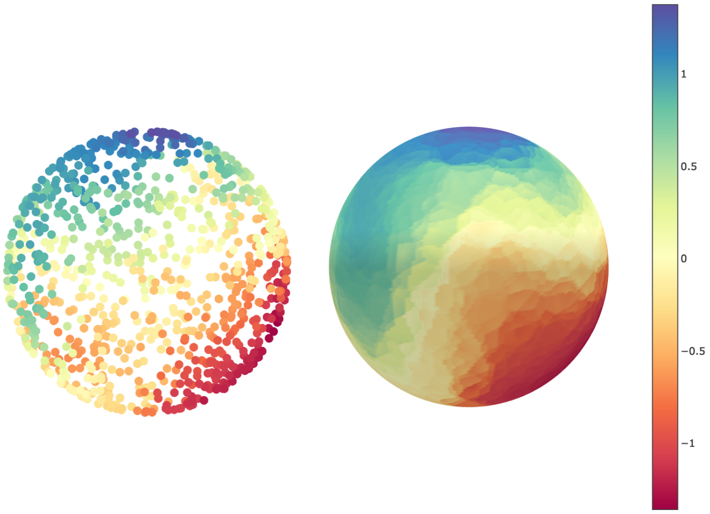

Figure 5: Draw from the discrete graph prior on the left, and the corresponding representation visualized using a 4-nearest-neighbors interpolation on the right. Parameters are $\varepsilon = 2n^{-1/4}$, $n = 1000$.

recovery of the initial condition with $t = 0.3$ is unsuccessful: the data does not contain enough information to accurately reconstruct $u^\dagger$. Figure 7 shows graph-posterior means computed in the regime of the first row of Figure 6 using the three graphs in Figure 1. Note that the spectra of the associated graph-Laplacians is represented in Figure 2. It is clear that inappropriate choice of ε leads to poor approximation of the continuum posterior, and here also to poor recovery of the initial condition $u^\dagger$. This is unsurprising in view of the dramatic effect of the choice of ε in the approximation properties of the spectrum of the spherical Laplacian, as shown in Figure 2. Note that while the numerical results are outside the asymptotic regime ($n = 1000$ throughout), they illustrate the role of ε. Theorem 3 establishes appropriate scalings for successful graph-learning in the large n asymptotic setting.

5.3. Algorithmic Scalability

It is important to stress that the large n robust performance of pCN methods established in this paper hinges on the existence of a continuum limit for the measures $\boldsymbol{\mu_n}$. Indeed, the fact that the limit posterior $\boldsymbol{\mu}$ over infinite dimensional functions can be written as a change of measure from the limit prior $\boldsymbol{\pi}$ has been rigorously shown to be equivalent to the limit learning problem having *finite* intrinsic dimension (Agapiou et al., 2017). In such a case, a key principle for the robust large n sampling of the measures $\boldsymbol{\mu_n}$ is to exploit the existence of a limit density, and use some variant of the dominating measure to obtain proposal samples. It has been established —and we do so here in the context of graph-based methods— that careful implementation of this principle leads to robust MCMC and importance sampling methodologies (Hairer et al., 2014; Agapiou et al., 2017).

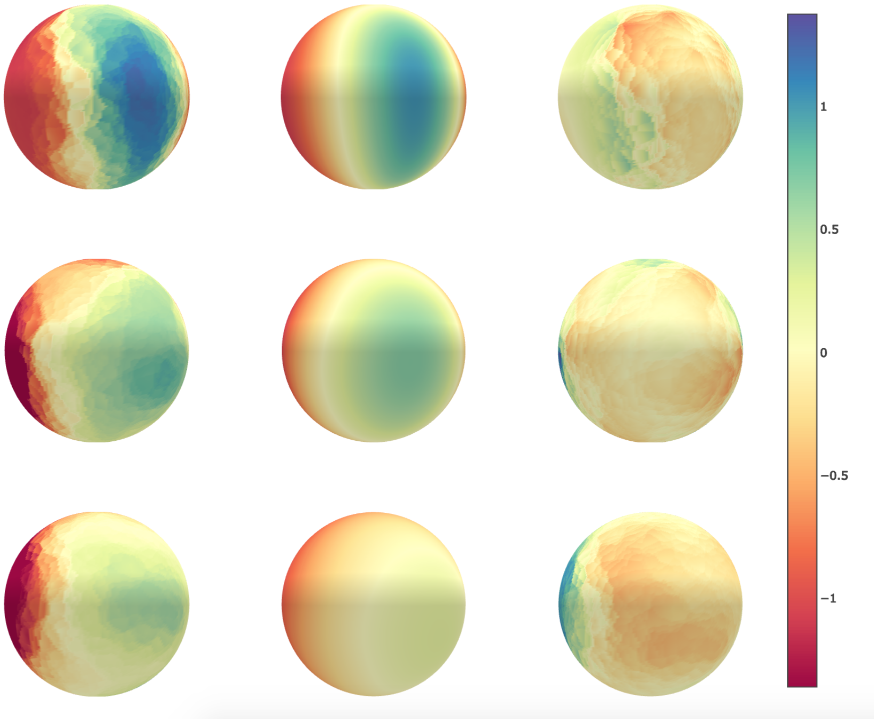

Figure 6: Means of the discrete and continuum posterior distributions are compared; the plots $P_{i,j}$ are arranged such that $P_{i,1}$ are graph-posterior means, $P_{i,2}$ are continuum posterior means, and $P_{i,3}$ are the differences row-wise. $P_{1,j},\ P_{2,j},\ P_{3,j}$ differ in the choice of the time parameter. They are, from the top, $t = [0,\ 0.1,\ 0.3]$.

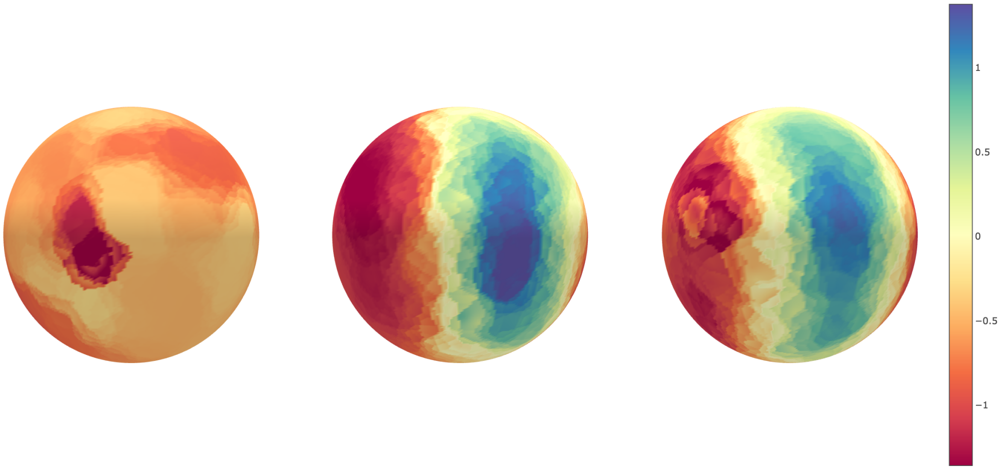

Figure 7: Graph-posterior means computed with the graph-pCN algorithm. All parameters of the learning problem are fixed to $t = 0,\ \sigma = 0.1,\ n = 1000,$ and $p = 200$. The three plots show three choices of graph connectivities $\varepsilon = [1,\ 2,\ 3] \times n^{-1/4}$ as in Figure 1.

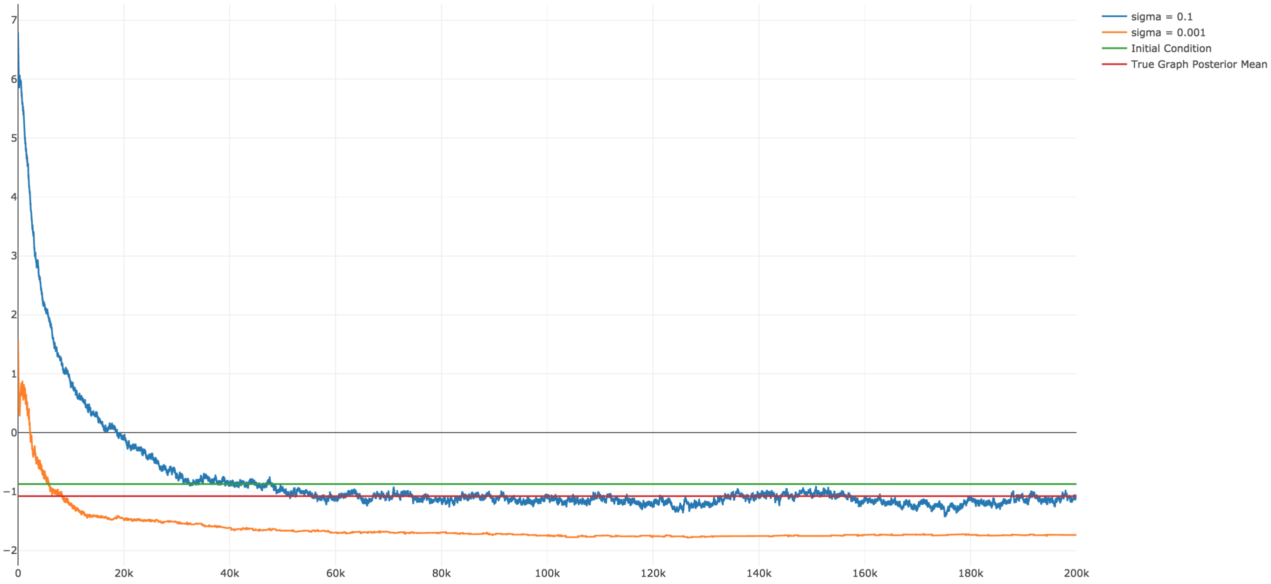

Figure 8: Effect of the parameter σ on graph-pCN algorithm. When σ is prohibitively small, here $\sigma = 0.001$, the chain fails to mix rapidly. With more noise, here $\sigma = 0.1$, the chain mixes rapidly.

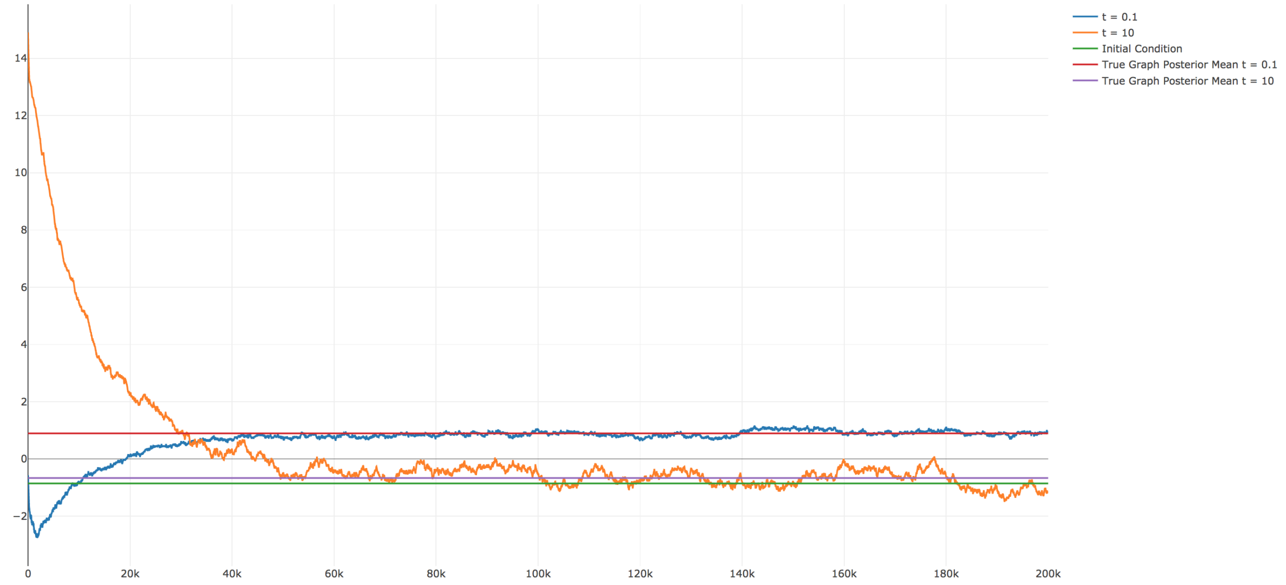

Figure 9: Shown here is the graph-pCN's chain mixing and converging for different values of the parameter t. Other parameter values for both chains are the same; note that the variation from $t = 0.1$ to $t = 10$ does not significantly affect the characteristics of the chain.

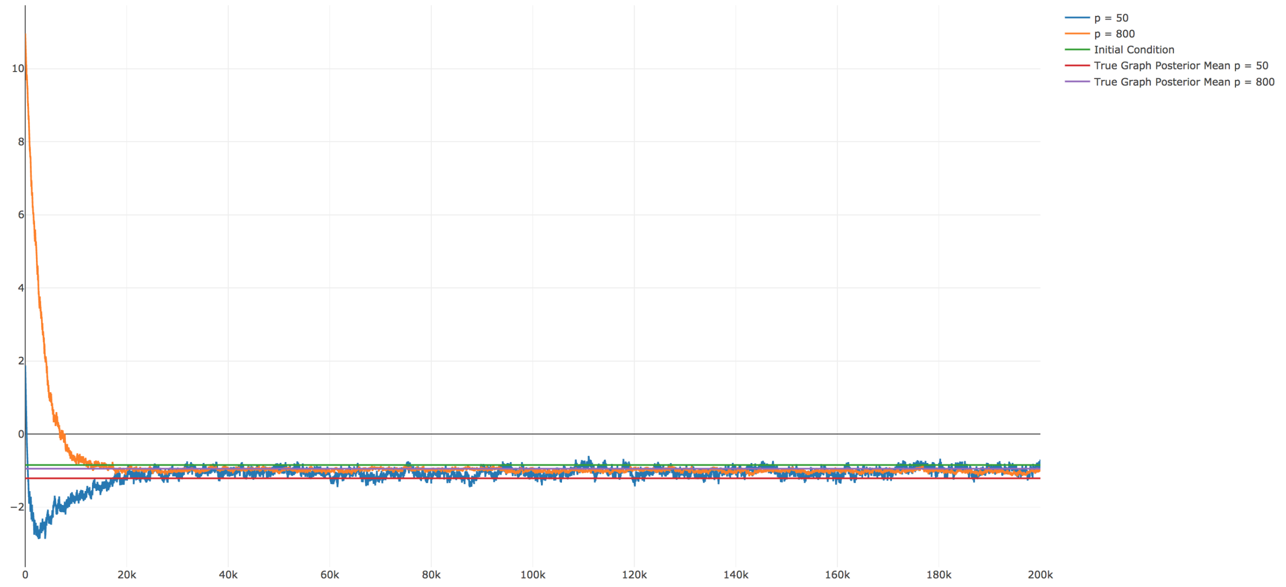

Figure 10: The above chart shows how increasing the value of the parameter p reduces the variance of the chain. Again, the chains above are both from the graph-pCN algorithm, and all other parameters are chosen so that the algorithm performs optimally.

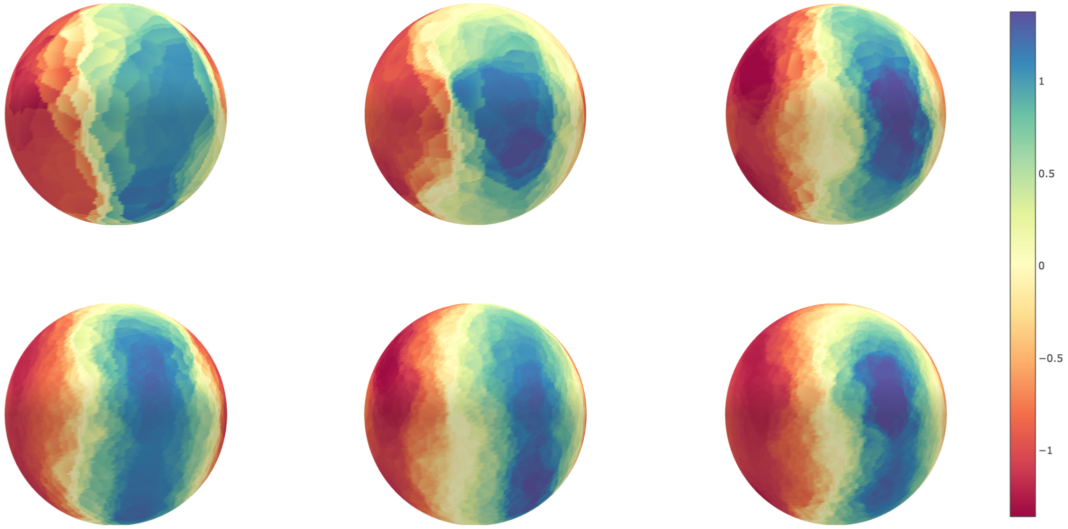

Figure 11: Graph pCN's robustness with respect to a changing value of n. In all plots $P_{i,j}$ above, $p = 200$, $t = 0.1$, $\sigma = 0.1$, and $\varepsilon = 2n^{-1/4}$. The plots are arranged such that $n = [300, 600, 900]$ for $P_{1,j}$ and $n = [1200, 1500, 2000]$ for $P_{2,j}$. The average acceptance probability remains constant with fixed β, as shown in Table 1.

A further point to note is that —even though from a theoretical and applied viewpoint it is clearly desirable that the data is informative— computational challenges in Bayesian settings often arise when the data is highly informative. This is also the case in the context of importance sampling and particle filters (Agapiou et al., 2017; Sanz-Alonso, 2018), where certain notion of distance between prior and proposal characterizes the algorithmic complexity. In the context of the pCN MCMC algorithms, if Φ is constant, the algorithm has acceptance probability 1. On the other hand, large Lipschitz constant of Φ (which translates to a posterior that is far from the prior) leads to small spectral gap. Indeed, tracking the spectral gap of pCN in terms of model parameters via the understanding of Lipschitz constants is in principle possible, and will be the subject of further work. In particular, small observation noise σ leads to deterioration of the pCN performance, see Figure 8. This issue may be alleviated by the use of the generalized version of pCN introduced in Rudolf and Sprungk (2015). Figures 9 and 10 investigate the role of the parameters t and p. All these figures show the posterior mean at one of the inputs, and the true graph posterior means have been computed with the formulae in the appendix.

Table 1 shows the large n robustness of pCN methods, while table 2 exhibits its deterioration in the fully supervised case $n = p$. The tables show the average acceptance probability with model parameters $\beta = 0.01$, $p = 200$, $\varepsilon_n = 2n^{-1/4}$ for the semi-supervised setting, and same parameters but with $p = n$ for the fully supervised case. The corresponding graph-posterior means are shown in Figure 11.

Table 1: Average acceptance probability for the graph pCN in the semi-supervised setting with constant data-set of size $p = 200$ and increasing number of unlabeled data.

n	300	600	900	1200	1500	2000
Acceptance Probability	0.230	0.245	0.237	0.249	0.236	0.239

Table 2: Deterioration of the average acceptance probability in a fully-supervised setting with $n = p$. The parameter β was held constant at $\beta = 0.01$. Additionally, $\varepsilon = 2n^{-1/4}$ and $t = 0$.

$n = p$	300	600	900	1200	1500	2000
Acceptance Probability	0.4536	0.3144	0.2360	0.1924	0.1644	0.1100

6. Acknowledgements

The work of NGT and DSA was supported by the NSF Grant DMS-1912818/1912802. ZK was funded by the NSF grant #1148284 IDyaS; TS would like to thank the Brown Division of Applied Mathematics for providing funds for the research. The authors are thankful to Dejan Slepčev for a careful reading of a first version of this manuscript.

References

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, 32(3):405–431, 2017.

S. R. Arridge, J. P. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen. Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Problems*, 22(1):175, 2006.

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017. URL `http://jmlr.org/papers/v18/15-205.html`.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

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

M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *COLT*, volume 3559, pages 486–500. Springer, 2005.

M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. *Advances in Neural Information Processing Systems (NIPS)*, 19:129, 2007.

M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

A. L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):568–595, 2018.

A. Beskos, G. O. Roberts, A. M. Stuart, and J. Voss. MCMC methods for diffusion bridges. *Stochastics and Dynamics*, 8(03):319–350, 2008.

A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. 2001.

D. Burago, S. Ivanov, and Y. Kurylev. A graph discretization of the Laplace-Beltrami operator. *J. Spectr. Theory*, 4:675–714, 2014.

S. L. Cotter, M. Dashti, J. C. Robinson, and A. M. Stuart. Bayesian inverse problems for functions and applications to fluid mechanics. *Inverse problems*, 25(11):115008, 2009.

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

T. Cui, Y. M. Marzouk, and K. E. Willcox. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering*, 102(5):966–990, 2015.

M. Dashti and A. M. Stuart. The bayesian approach to inverse problems. Handbook of Uncertainty Quantification.

D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10): 5591–5596, 2003.

A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan. Asymptotic behavior of l_p-based Laplacian regularization in semi-supervised learning. In *Conference on Learning Theory*, pages 879–906, 2016.

T. Gao, S. Z. Kovalsky, and I. Daubechies. Gaussian process landmarking on manifolds. *SIAM Journal on Mathematics of Data Science*, 1(1):208–236, 2019.

N. García Trillos and R. Murray. A new analytical approach to consistency and overfitting in regularized empirical risk minimization. *European Journal of Applied Mathematics*, pages 1–36, 2017.

N. García Trillos and D. Sanz-Alonso. The Bayesian formulation and well-posedness of fractional elliptic inverse problems. *Inverse Problems*, 33(6):065006, 2017.

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

N. García Trillos and D. Sanz-Alonso. The Bayesian update: variational formulations and gradient flows. *Bayesian Analysis*, 2018b.

N. García Trillos and D. Slepčev. On the rate of convergence of empirical measures in ∞-transportation distance. *Canadian Journal of Mathematics*, 67:1358–1383, 2014.

N. García Trillos and D. Slepčev. Continuum limit of total variation on point clouds. *Archive for rational mechanics and analysis*, 220(1):193–241, 2016a.

N. García Trillos and D. Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 2016b.

N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev. Error estimates for spectral convergence of the graph laplacian on random geometric graphs towards the laplace–beltrami operator. *arXiv preprint arXiv:1801.10108*, 2018.

N. García Trillos, D. Sanz-Alonso, and R. Yang. Local regularization of noisy point clouds: Improved global geometric estimates and data analysis. *Journal of Machine Learning Research*, 20(136):1–37, 2019. URL `http://jmlr.org/papers/v20/19-261.html`.

C. J. Geyer and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.

E. Giné and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results. In *High dimensional probability*, volume 51 of *IMS Lecture Notes Monogr. Ser.*, pages 238–259. Inst. Math. Statist., Beachwood, OH, 2006. doi: 10.1214/074921706000000888. URL `http://dx.doi.org/10.1214/074921706000000888`.

M. Hairer, J. C. Mattingly, and M. Scheutzow. Asymptotic coupling and a general form of Harrisâ theorem with applications to stochastic delay equations. *Probability theory and related fields*, 149(1):223–259, 2011.

M. Hairer, A. M. Stuart, and S. J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24(6):2455–2490, 2014.

J. Harlim, D. Sanz-Alonso, and R. Yang. Kernel methods for Bayesian elliptic inverse problems on manifolds. *arXiv preprint arXiv:1910.10669*, 2019.

J. Hartog and H. van Zanten. Nonparametric Bayesian label prediction on a graph. *arXiv preprint arXiv:1612.01930*, 2016.

M. Hein. Uniform convergence of adaptive graph-based regularization. In G. Lugosi and H. U. Simon, editors, *Proc. of the 19th Annual Conference on Learning Theory (COLT)*, pages 50–64. Springer, 2006.

M. Hein, J-Y Audibert, and U. Von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *Journal of Machine Learning Research*, 8(Jun):1325–1368, 2007.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

C. Kipnis and S. R. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.

F. Liang, S. Mukherjee, and M. West. The use of unlabeled data in predictive modeling. *Statistical Science*, pages 189–205, 2007.

M. Maier, U. Von Luxburg, and M. Hein. Influence of graph construction on graph-based clustering measures. In *Advances in neural information processing systems*, pages 1025–1032, 2009.

Y. M. Marzouk, H. N. Najm, and L. A. Rahn. Stochastic spectral methods for efficient Bayesian solution of inverse problems. *Journal of Computational Physics*, 224(2):560–586, 2007.

P. J. Olver. *Introduction to Partial Differential Equations*. Springer Science & Business Media, 2013.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*, volume 1. MIT press Cambridge, 2006.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

D. Rudolf and B. Sprungk. On a generalization of the Preconditioned Crank–Nicolson Metropolis algorithm. *Foundations of Computational Mathematics*, pages 1–35, 2015.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

D. Sanz-Alonso. Importance sampling and necessary sample size: an information theory approach. *SIAM/ASA Journal on Uncertainty Quantification*, 6(2):867–879, 2018.

Z. Shi. Convergence of Laplacian spectra from random samples. arXiv preprint arXiv:1507.00151, 2015.

V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 824–831. ACM, 2005.

A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

A. Singer and H-T Wu. Spectral convergence of the connection Laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1):58–123, 2017.

D. Slepčev and M. Thorpe. Analysis of p-Laplacian regularization in semi-supervised learning. *arXiv preprint arXiv:1707.06213*, 2017.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 2012.

A. M. Stuart and A. Teckentrup. Posterior consistency for Gaussian process approximations of Bayesian posterior distributions. *Mathematics of Computation*, 2017.

J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

D. Ting, L. Huang, and M. I. Jordan. An analysis of the convergence of graph Laplacians. In *Proc. of the 27th Int. Conference on Machine Learning (ICML)*, 2010.

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

D. Xiu. *Numerical methods for stochastic computations: a spectral method approach.* Princeton University Press, 2010.

D. Zhou and B. Schölkopf. Regularization on discrete spaces. In *Joint Pattern Recognition Symposium*, pages 361–368. Springer, 2005.

X. Zhu. Semi-supervised learning literature survey. 2005.

X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

Appendix A. Benchmark Formulae

Here we exploit the linearity of the forward and observation maps to compute, under the Gaussian observation noise model, the mean and covariance of the Gaussian graph and continuum posteriors. These formulae could be useful in understanding the approximation of continuum posteriors by graph posteriors, and to provide benchmarks for posteriors computed with MCMC methods. For the derivations we use the covariance function representation of Gaussian measures and the theory of Gaussian process regression in Rasmussen and Williams (2006). Throughout we assume that s is large enough so that the formulae below are well-defined.

We start with the continuum case. Set $v := \mathcal{F}u$. The prior (6) on u induces a prior on $v \sim GP\big(0, c_v(x,\tilde{x})\big)$, where

$$c_v(x,\tilde{x}) = \sum_{i=1}^{\infty} e^{-2\lambda_i t}(\alpha+\lambda_i)^{-s/2}\psi_i(x)\psi_i(\tilde{x}). \tag{26}$$

Then, we have a regression problem for v given data $y = [y_1, \dots, y_p]'$

$$y_i = v(\mathbf{x}_i) + \eta_i, \quad \eta_i \sim N(0,\gamma^2)$$

in the form of Rasmussen and Williams (2006). The posterior distribution of $v|y$ is thus given by a Gaussian process $GP\big(m_{v|y}(x), c_{v|y}(x,\tilde{x})\big)$, with

$$\begin{aligned} m_{v|y}(x) &= c_v(x,X)'\big(c_v(X,X)+\gamma^2 I\big)^{-1} y, \\ c_{v|y}(x,\tilde{x}) &= c_v(x,\tilde{x}) - c_v(x,X)'\big(c_v(X,X)+\gamma^2 I\big)^{-1} c_v(\tilde{x},X), \end{aligned}$$

where we use the following notations:

$$c_v(x, X) := [c_v(x, \mathbf{x}_1), \ldots, c_v(x, \mathbf{x}_p)]' \in \mathbb{R}^p,$$
$$c_v(X, X) := \big(c_v(\mathbf{x}_i, \mathbf{x}_j)\big)_{1 \le i,j \le p} \in \mathbb{R}^{p \times p}.$$

Now the posterior of interest $\boldsymbol{\mu}$ on u given y can be recovered by running the heat equation backwards. Namely, we have that $\boldsymbol{\mu} = GP\big(m_{u|y}(x), c_{u|y}(x, \tilde{x})\big)$ with

$$\begin{aligned} m_{u|y}(x) &= c_w(x, X)'(c_v(X, X) + \gamma^2 I)^{-1} y, \\ c_{u|y}(x, \tilde{x}) &= c_u(x, \tilde{x}) - c_w(x, X)'\big(c_v(X, X) + \gamma^2 I\big)^{-1} c_w(\tilde{x}, X), \end{aligned} \tag{27}$$

where $c_w(x, X)$ is a vector made of evaluations of the covariance function of $w := \mathcal{F}^{1/2} u$ at the test and training points. Precisely, its j-th entry is given by

$$c_w(x, X)_j = \sum_{i=1}^{\infty} e^{-\lambda_i t} (\alpha + \lambda_i)^{-s/2} \psi_i(x) \psi_i(\mathbf{x}_j). \tag{28}$$

There are several points to note about equation (27). First, the predictive mean is a linear function of the data y, hence a linear predictor. It is indeed the best linear predictor in a mean-squared error sense (Stein, 2012). Second, since $c_v(X, X) + \gamma^2 I$ is positive definite, $c_{u|y}(x, \tilde{x}) \le c_u(x, \tilde{x})$; thus, conditioning reduces the uncertainty. Moreover, in the limit of noiseless observations ($\gamma = 0$) and $t = 0$ we recover that $c_{u|y}(\mathbf{x}_i, \mathbf{x}_j) = 0$ in the training points. However, even with noiseless observations this is not true if $t > 0$. Finally, note the well-known fact that the the posterior covariance $c_{u|y}$ does not depend on the observed data y.

Formulae in the discrete setting can be obtained in a similar way, and we omit the details. Plugging in the data y from the continuum setting, we deduce that

$$\boldsymbol{\mu_n} = N\big(m_{u_n|y}(\mathbf{x}_k), c_{u_n|y}(\mathbf{x}_k, \mathbf{x}_l)\big),$$

with

$$\begin{aligned} m_{u_n|y}(\mathbf{x}_k) &= c_{w_n}(x, X)'(c_{v_n}(X, X) + \gamma^2 I)^{-1} y, \\ c_{u_n|y}(\mathbf{x}_k, \mathbf{x}_l) &= c_{u_n}(\mathbf{x}_k, \mathbf{x}_l) - c_{w_n}(\mathbf{x}_k, X)'\big(c_{v_n}(X, X) + \gamma^2 I\big)^{-1} c_{w_n}(\mathbf{x}_l, X). \end{aligned} \tag{29}$$

In the above equations, all objects indexed by n constitute straightforward analogues of objects in the continuum, constructed using the graph spectrum rather than the continuum one.

Appendix B. The TL^2 and $\mathcal{P}(TL^2)$ Spaces

Let us recall the definition of the TL^2 space. First, we define the set

$$TL^2 := \big\{(\theta, f) \; : \; \theta \in \mathcal{P}(\mathcal{M}), \, f \in L^p(\mathcal{M}, \theta)\big\}.$$

Then, for arbitrary elements (θ_1, f_1) and (θ_2, f_2) in TL^2 we define, following García Trillos and Slepčev (2016a),

$$d_{TL^2}\big((\theta_1, f_1), (\theta_2, f_2)\big) := \inf_{\omega \in \Gamma(\theta_1, \theta_2)} \left(\iint_{\mathcal{M}\times\mathcal{M}} \Big(d_{\mathcal{M}}(x,y)^2 + |f_1(x) - f_2(y)|^2 \Big) d\omega(x,y) \right)^{1/2}, \tag{30}$$

where $\Gamma(\theta_1, \theta_2)$ is the set of Borel probability measures on $\mathcal{M} \times \mathcal{M}$ with marginal θ_1 on the first factor and θ_2 on the second one. It was shown in García Trillos and Slepčev (2016a) that d_{TL^2} defines a distance in TL^2.

The TL^2 space allows us to make sense of a sequence $u_n \in L^2(\gamma_n)$ converging towards an element $u \in L^2(\gamma)$. Indeed, with a slight abuse of notation, we say that a sequence $u_n \in L^2(\gamma_n)$ converges in TL^2 towards $u \in L^2(\gamma)$, written

$$u_n \overset{TL^2}{\longrightarrow} u,$$

if $d_{TL^2}\big((u_n, \gamma_n), (u, \gamma)\big) \to 0$. A characterization of convergence in TL^2 in terms of composition with transport maps can be found in Proposition 3.12 in García Trillos and Slepčev (2016a).

As noted in García Trillos and Slepčev (2016a), (TL^2, d_{TL^2}) is not a complete metric space. Its completion however, denoted $\overline{TL^2}$, can be identified with the space $\mathcal{P}_2(\mathcal{M} \times \mathbb{R})$ of Borel probability measures on the product space $\mathcal{M} \times \mathbb{R}$ with finite second moments, endowed with the Wasserstein distance. The space $\overline{TL^2}$ is a Polish space.

Having introduced the metric space TL^2 we can now define $\mathcal{P}(TL^2)$ to be the space of Borel probability measures on TL^2 endowed with the weak convergence of probability measures. If $\theta \in \mathcal{P}(\mathcal{M})$ and $\boldsymbol{\nu} \in \mathcal{P}(L^2(\theta))$, it is possible to think of $\boldsymbol{\nu}$ as an element in $\mathcal{P}(TL^2)$. Indeed, the canonical inclusion

$$\mathcal{I}_\theta : f \in L^2(\theta) \longmapsto (\theta, f) \in TL^2$$

induces the canonical inclusion

$$\mathcal{I}_{\theta\sharp} : \mathcal{P}(L^2(\theta)) \hookrightarrow \mathcal{P}(TL^2),$$

where $\mathcal{I}_{\theta\sharp}$ is the push-forward via $\mathcal{I}_\theta$. Notice that $\mathcal{I}_\theta$ is a continuous map. In the sequel we may drop the explicit mention to $\mathcal{I}$ whenever no confusion arises from doing so.

The above observation motivates the following definition.

Definition 12 *For $\boldsymbol{\nu_n} \in \mathcal{P}\big(L^2(\gamma_n)\big)$, $n \in \mathbb{N}$, and $\boldsymbol{\nu} \in \mathcal{P}\big(L^2(\gamma)\big)$ we say that $\{\boldsymbol{\nu_n}\}_{n\in\mathbb{N}}$ converges to $\boldsymbol{\nu}$, written*

$$\boldsymbol{\nu_n} \overset{\mathcal{P}(TL^2)}{\longrightarrow} \boldsymbol{\nu},$$

if $\{\mathcal{I}_{\gamma_n\sharp}\boldsymbol{\nu_n}\}_{n\in\mathbb{N}}$ converges weakly to $\mathcal{I}_{\gamma\sharp}\boldsymbol{\nu}$ in $\mathcal{P}(TL^2)$.

This is the notion of convergence of discrete to continuum posteriors that we use in this paper. The space $\mathcal{P}(TL^2)$ was introduced in García Trillos and Sanz-Alonso (2018a).

Appendix C. Proof of Theorem 3

We want to show that

$$\mathcal{I}_{n\sharp}\boldsymbol{\mu_n} \to_{\mathcal{P}(L^2(\gamma))} \boldsymbol{\mu}, \qquad \text{as } n \to \infty. \tag{31}$$

Step 0: The proof of Theorem 4.1 in García Trillos and Sanz-Alonso (2018a) shows that

$$\boldsymbol{\pi}_n \to_{\mathcal{P}(TL^2)} \boldsymbol{\pi}, \qquad \text{as } n \to \infty,$$

under the assumptions of Theorem 3 (in particular removing the upper bound assumption on ε_n from Theorems 4.1 and 4.4 in García Trillos and Sanz-Alonso (2018a)). Likewise the proof of Theorem 4.4 in García Trillos and Sanz-Alonso (2018a) establishes the Γ-convergence of the energies

$$J_n(\boldsymbol{\nu_n}) := D_{\mathrm{KL}}(\boldsymbol{\nu_n}\|\boldsymbol{\pi_n}) + \int_{L^2(\gamma_n)} \phi_n(u_n;y)d\boldsymbol{\nu_n}(u_n), \quad \boldsymbol{\mu_n} \in \mathcal{P}(L^2(\gamma_n)),$$

towards the energy

$$J(\boldsymbol{\nu}) = D_{\mathrm{KL}}(\boldsymbol{\nu}\|\boldsymbol{\pi}) + \int_{L^2(\gamma)} \phi(u;y)d\boldsymbol{\nu}(u), \quad \boldsymbol{\nu} \in \mathcal{P}(L^2(\gamma))$$

in the $\mathcal{P}(TL^2)$-sense, under the assumptions of Theorem 3. In particular,

$$\boldsymbol{\mu_n} \to_{\mathcal{P}(TL^2)} \boldsymbol{\mu}, \quad n \to \infty,$$

because $\boldsymbol{\mu_n}$ is the minimizer of J_n and $\boldsymbol{\mu}$ is the minimizer of J (see the variational characterization of posterior distributions in García Trillos and Sanz-Alonso (2018b)).

Step 1: We claim that $\{\mathcal{I}_{n\sharp}\boldsymbol{\mu_n}\}_{n\in\mathbb{N}}$ is pre-compact with respect to the weak convergence of probability measures on $L^2(\gamma)$. By Lemma 5.1 in García Trillos and Sanz-Alonso (2018a) it is enough to show that

(i) $\sup_{n\in\mathbb{N}} D_{\mathrm{KL}}(\mathcal{I}_{n\sharp}\boldsymbol{\mu_n}\|\mathcal{I}_{n\sharp}\boldsymbol{\pi_n}) < +\infty$; and

(ii) $\mathcal{I}_{n\sharp}\boldsymbol{\pi_n} \to_{\mathcal{P}(L^2(\gamma))} \boldsymbol{\pi}$.

Let us start with (i). Step 0 implies that

$$\lim_{n\to\infty} \min_{\boldsymbol{\nu_n}} J_n(\boldsymbol{\nu_n}) = \min_{\boldsymbol{\nu}} J(\boldsymbol{\nu}) < +\infty.$$

Given that $\boldsymbol{\mu_n}$ is the minimizer of J_n and $\boldsymbol{\mu}$ is the minimizer of J, it follows that

$$\lim_{n\to\infty} J_n(\boldsymbol{\mu_n}) = J(\boldsymbol{\mu}) < +\infty.$$

Combining the previous fact with the chain of inequalities

$$D_{\mathrm{KL}}(\mathcal{I}_{n\sharp}\boldsymbol{\mu_n}\|\mathcal{I}_{n\sharp}\boldsymbol{\pi_n}) \le D_{\mathrm{KL}}(\boldsymbol{\mu_n}\|\boldsymbol{\pi_n}) \le J_n(\boldsymbol{\mu_n})$$

gives (i).

We now show (ii). Consider an orthonormal basis of eigenvectors $\{\psi_1^n, \dots, \psi_i^n\}$ of $\Delta_{\mathcal{M}_n}$ and an orthonormal basis $\{\psi_1, \dots, \psi_n, \dots\}$ of eigenfunctions of $\Delta_{\mathcal{M}}$. By the results in

García Trillos and Slepčev (2016b) we can assume without the loss of generality that, for all $j \in \mathbb{N}$,

$$\psi_j^n \to_{TL^2} \psi_j, \text{ as } n \to \infty.$$

Let $(\tilde{\Omega}, \tilde{F}, \tilde{\mathbb{P}})$ be a probability space supporting i.i.d. random variables $\{\xi_i\}_{i \in \mathbb{N}}$ with $\xi_i \sim N(0,1)$ and consider

$$X_n = \sum_{i=1}^{k_n} (\alpha + \lambda_i^n)^{-s/4} \xi_i \psi_i^n, \quad X = \sum_{i=1}^{\infty} (\alpha + \lambda_i)^{-s/4} \xi_i \psi_i,$$

where, recall, k_n is the truncation level of the prior $\boldsymbol{\pi_n}$. Notice that $X_n \sim \boldsymbol{\pi_n}$, $X \sim \boldsymbol{\pi}$ and $\mathcal{I}_n(X_n)$ is distributed according to $\mathcal{I}_{n\sharp}\boldsymbol{\pi_n}$. For any fixed $i = 1, \dots, k_n$ it follows from the first part of the proof of Theorem 1.10 in García Trillos et al. (2018) that

$$\|\mathcal{I}_n(\psi_i^n)\|_{L^2(\gamma)} \leq \|\mathcal{I}_n(\psi_i^n) - \psi_i\|_{L^2(\gamma)} + \|\psi_i\|_{L^2(\gamma)} \leq C, \tag{32}$$

where C is a constant independent of $i = 1, \dots, k_n$ and n. It then follows that for every $l \in \mathbb{N}$,

$$\begin{aligned}
\|\mathcal{I}_n(X_n) - X\|_{L^2(\gamma)} \leq & \left\| \sum_{i=1}^{l} (\alpha + \lambda_i^n)^{-s/4} \xi_i \mathcal{I}_n(\psi_i^n) - \sum_{i=1}^{l} (\alpha + \lambda_i)^{-s/4} \xi_i \psi_i \right\|_{L^2(\gamma)} \\
& + \sum_{i=l}^{k_n} (\alpha + \lambda_i^n)^{-s/4} |\xi_i| \|\mathcal{I}_n(\psi_i^n)\|_{L^2(\gamma)} + \sum_{i=l}^{\infty} (\alpha + \lambda_i)^{-s/4} |\xi_i| \|\psi_i\|_{L^2(\gamma)} \\
\leq & \left\| \sum_{i=1}^{l} (\alpha + \lambda_i^n)^{-s/4} \xi_i \mathcal{I}_n(\psi_i^n) - \sum_{i=1}^{l} (\alpha + \lambda_i)^{-s/4} \xi_i \psi_i \right\|_{L^2(\gamma)} + C \sum_{i=l}^{\infty} (\alpha + \lambda_i)^{-s/4} |\xi_i|,
\end{aligned}$$

where C is a constant that does not depend on n; we have used the bounds (32) on $\|\mathcal{I}_n(\psi_i^n)\|_{L^2(\gamma)}$ and the bounds (24) for λ_i^n in terms of λ_i for $i = 1, \dots, k_n$. We can then take expectations and lim sups in both sides of the above inequality and use Theorem 1.10 in García Trillos et al. (2018) to conclude that

$$\limsup_{n \to \infty} \mathbb{E}\left(\|\mathcal{I}_n(X_n) - X\|_{L^2(\gamma)} \right) \leq C \sum_{i=l}^{\infty} (\alpha + \lambda_i)^{-s/4}.$$

Since the above is true for every l and the series is convergent, (ii) follows.

An application of Lemma 5.1 in García Trillos and Sanz-Alonso (2018a) allows us to deduce that $\{\mathcal{I}_{n\sharp}\boldsymbol{\mu_n}\}_{n \in \mathbb{N}} \subseteq \mathcal{P}(L^2(\gamma))$ is pre-compact and, moreover, that each of its cluster points is a measure that is absolutely continuous with respect to $\boldsymbol{\pi}$. We can then assume without the loss of generality that, for some $\tilde{\boldsymbol{\mu}} \in \mathcal{P}(L^2(\gamma))$,

$$\mathcal{I}_{n\sharp}\boldsymbol{\mu_n} \to_{\mathcal{P}(L^2(\gamma))} \tilde{\boldsymbol{\mu}}, \quad \text{as } n \to \infty.$$

Step 2: To show (31) it is then enough to prove that the finite dimensional projections of $\tilde{\boldsymbol{\mu}}$ coincide with those of $\boldsymbol{\mu}$. More precisely, we identify $u \in L^2(\gamma)$ with the infinite vector

$(u_1, u_2, \dots)$ denoting the coefficients of u in the basis $\{\psi_1, \psi_2, \dots\}$ and define $\mathrm{Proj}_j(u) := \sum_{i=1}^{j} u_i \psi_i$; we need to show that for arbitrary $j \in N$ we have

$$\mathrm{Proj}_{j\sharp}\, \tilde{\boldsymbol{\mu}} = \mathrm{Proj}_{j\sharp}\, \boldsymbol{\mu}.$$

From Step 0 and Skorohod's theorem, we know there exists a probability space $(\tilde{\Omega}, \tilde{F}, \tilde{\mathbb{P}})$ supporting random variables $\{X_n^y\}_{n\in\mathbb{N}}$ and X^y with $X_n^y \sim \boldsymbol{\mu_n}$ and $X^y \sim \boldsymbol{\mu}$ and for which $X_n^y \to_{TL^2} X^y$ for $\tilde{\mathbb{P}}$-a.e. $\tilde{\omega} \in \tilde{\Omega}$. We can then write

$$X_n^y = \sum_{i=1}^{k_n} a_i^n \psi_i^n, \quad X^y = \sum_{i=1}^{\infty} a_i \psi_i,$$

for some random variables a_i^n and a_i. Notice that the continuity of inner products with respect to TL^2-convergence (see Proposition 2.6 in García Trillos and Slepčev (2016b)) implies that

$$\lim_{n\to\infty} a_i^n = a_i, \quad \tilde{\mathbb{P}}\text{-a.e.}$$

Now, for every fixed $l \geq j$ we can write

$$\mathrm{Proj}_j(\mathcal{I}_n(X_n^y)) = \sum_{i=1}^{l} a_i^n \,\mathrm{Proj}_j(\mathcal{I}_n(\psi_i^n)) + \sum_{i=l+1}^{k_n} a_i^n \,\mathrm{Proj}_j(\mathcal{I}_n(\psi_i^n)). \tag{33}$$

The left hand side of the above expression is seen to converge weakly towards $\mathrm{Proj}_{j\sharp}\, \tilde{\boldsymbol{\mu}}$ because $\mathcal{I}_n(X_n^y) \sim \mathcal{I}_{n\sharp}\boldsymbol{\mu_n}$, $\mathcal{I}_{n\sharp}\boldsymbol{\mu_n} \to_{\mathcal{P}(L^2(\gamma))} \tilde{\boldsymbol{\mu}}$, and because Proj_j is continuous. On the other hand, the first term on the right hand side is seen to converge $\tilde{\mathbb{P}}$-a.e. towards $\sum_{i=1}^{j} a_i \,\mathrm{Proj}_j(\psi_i) = \sum_{i=1}^{j} a_i \psi_i$ because

$$\mathcal{I}_n(\psi_i^n) \to_{L^2(\gamma)} \psi_i, \qquad \text{as } n \to \infty,$$

which follows from Theorem 1.10 in García Trillos et al. (2018) (it is at this stage that we need the extra technical condition on ε_n); in particular this term converges weakly towards $\mathrm{Proj}_{j\sharp}\, \boldsymbol{\mu}$. To show $\mathrm{Proj}_{j\sharp}\, \tilde{\boldsymbol{\mu}} = \mathrm{Proj}_{j\sharp}\, \boldsymbol{\mu}$ it is then enough, by Slutsky's theorem, to prove that $\|\sum_{i=l+1}^{k_n} a_i^n \,\mathrm{Proj}_j(\mathcal{I}_n(\psi_i^n))\|_{L^2(\gamma)}$ converges in probability towards zero.

To see this, first notice that

$$\left\| \sum_{i=l+1}^{k_n} a_i^n \,\mathrm{Proj}_j(\mathcal{I}_n(\psi_i^n)) \right\|_{L^2(\gamma)} \leq C \sum_{i=l+1}^{k_n} |a_i^n|.$$

Fix $t > 0$. Observe that the expression

$$\limsup_{n\to\infty} \tilde{\mathbb{P}} \left(\left\| \sum_{i=l+1}^{k_n} a_i^n \,\mathrm{Proj}_j(\mathcal{I}_n(\psi_i^n)) \right\|_{L^2(\gamma)} > t \right)$$

is independent of l. Then,

$$q_j(t) := \limsup_{n\to\infty} \tilde{\mathbb{P}} \left(\left\| \sum_{i=l+1}^{k_n} a_i^n \operatorname{Proj}_j(\mathcal{I}_n(\psi_i^n)) \right\|_{L^2(\gamma)} > t \right)$$
$$\leq \limsup_{n\to\infty} \tilde{\mathbb{P}} \left(\sum_{i=l+1}^{k_n} |a_i^n| > \frac{t}{C} \right).$$

On the other hand, identifying the elements in the support of $\boldsymbol{\pi_n}$ with $\mathbb{R}^{k_n}$ (i.e. writing $u_n \in \operatorname{supp}(\boldsymbol{\pi_n})$ in the basis $\{\psi_1^n, \dots, \psi_{k_n}^n\}$) and letting $A_{n,t,l}$ be the set

$$A_{n,t,l} := \left\{ x \in \mathbb{R}^{k_n} \ : \ \sum_{i=l+1}^{k_n} |x_i| > \frac{t}{C} \right\},$$

we see that

$$\tilde{\mathbb{P}} \left(\sum_{i=l+1}^{k_n} |a_i^n| > \frac{t}{C} \right) = \boldsymbol{\mu_n}\left(A_{n,t,l}\right) = \frac{1}{Z_n} \int_{A_{n,t,l}} \exp(-\Phi_n(x;y)) d\boldsymbol{\pi_n}(x) \leq \frac{1}{Z_n} \boldsymbol{\pi_n}\left(A_{n,t,l}\right),$$

and hence

$$\limsup_{n\to\infty} \tilde{\mathbb{P}} \left(\sum_{i=l+1}^{k_n} |a_i^n| > \frac{t}{C} \right) \leq \frac{1}{Z} \boldsymbol{\pi} \left(\{ u \in L^2(\gamma) \ : \ \sum_{i=l+1}^{\infty} |u_i| > t/C \} \right).$$

In the above Z and Z_n are the normalization constants from (1) and (2) respectively.

Therefore,

$$q_j(t) \leq \frac{1}{Z} \boldsymbol{\pi} \left(\{ u \in L^2(\gamma) \ : \ \sum_{i=l+1}^{\infty} |u_i| > t/C \} \right).$$

Taking now the limit as $l \to \infty$ of the right hand side of the above expression, we deduce that $q_j(t) = 0$. Since this is true for arbitrary $t > 0$, we deduce that indeed $\|\sum_{i=l+1}^{k_n} a_i^n \operatorname{Proj}_j(\mathcal{I}_n(\psi_i^n))\|_{L^2(\gamma)}$ converges in probability towards zero and the proof is now complete.

Remark 13 *In the above proof we have used results from García Trillos et al. (2018) on Voronoi extensions, but it is clear that analogue results can be deduced for more general interpolation maps $\{\mathcal{I}_n\}_{n\in\mathbb{N}}$ as long as one can show the following:*

i) *(Uniform L^2-boundedness) There is a constant $C > 0$ such that $\|\mathcal{I}_n \psi_i^n\|_{L^2(\gamma)} \leq C$ for every $i = 1, \dots, k_n$ and for every n.*

ii) *(Consistency) For every $i \in \mathbb{N}$ we have $\mathcal{I}_n(\psi_i^n) \to_{L^2(\gamma)} \psi_i$.*

Appendix D. Proof of Theorem 7

The proof of Theorem 7 is based on the paper Hairer et al. (2014) which in turn makes use of the following weak form of Harris theorem from Hairer et al. (2011). We let $\mathcal{H}$ be a separable Hilbert space and for a distance like function $\tilde{d} : \mathcal{H} \times \mathcal{H} \to [0, \infty)$ define the associated Wasserstein distance (1-OT distance) on $\mathcal{P}(\mathcal{H})$

$$\tilde{d}(\mu, \nu) := \inf_{\theta \in \Gamma(\mu,\nu)} \int_{\mathcal{H}\times\mathcal{H}} \tilde{d}(u, w) d\theta(u, w), \quad \mu, \nu \in \mathcal{P}(\mathcal{H}), \tag{34}$$

where $\Gamma(\mu, \nu)$ denotes the set of couplings between μ and ν.

Theorem 14 (Weak Harris Theorem; Theorem 4.7 in Hairer et al. (2011)) *Let $\mathcal{H}$ be a separable Hilbert space and let P be a transition kernel for a discrete time Markov chain with state space $\mathcal{H}$ for which the following conditions are satisfied:*

i) *(Lyapunov functional) There exists a lower semi-continuous function $V : \mathcal{H} \to [0, \infty)$ such that*
$$PV(u) := \int_{\mathcal{H}} V(w)P(u, dw) \leq lV(u) + K, \quad \forall u \in \mathcal{H}, \tag{35}$$
where $K > 0$ and $0 < l < 1$ are some constants.

ii) *(d-contraction) There exist a distance like function $d : \mathcal{H} \times \mathcal{H} \to [0, 1]$ and a constant $\varrho \in (0, 1)$ such that, for all $u, w \in \mathcal{H}$ with $d(u, w) < 1$,*
$$d(u, w) \leq \varrho.$$

iii) *(d-smallness of level sets of V) For the distance like function d above, the functional V and the constant K in (35), there exists $\vartheta \in (0, 1)$ such that, for all u, w with $V(u), V(w) \leq 4K$,*
$$d(u, w) \leq \vartheta.$$

Then, the Markov chain P has a $\tilde{d}$-Wasserstein spectral gap where $\tilde{d}$ is the distance like function
$$\tilde{d}(u, w) = \sqrt{d(u, w)(1 + V(u) + V(w))}, \quad u, w \in \mathcal{H}.$$
More precisely, there exist $\lambda > 0$ and $C > 0$ such that

$$\tilde{d}(P^j\mu, P^j\nu) \leq C \exp(-\lambda j)\tilde{d}(\mu, \nu), \quad \forall \mu, \nu \in \mathcal{P}(\mathcal{H}), \quad \forall j \in \mathbb{N}.$$

Remark 15 *As remarked in Hairer et al. (2011), we highlight that the second hypothesis is an assumption that holds for points u, w with $d(u, w) < 1$ and that nothing is being stated about points for which $d(u, w) = 1$. The observation here is that even if one cannot deduce a Wasserstein spectral gap for the distance like function d, one can still obtain a Wasserstein spectral gap for the distance like function $\tilde{d}$.*

It is possible to quantify the constants λ and C in the conclusion of Theorem 14 in terms of the parameters l, K, ϱ, ϑ. Here, however, we are simply interested in pointing out how changing the parameters in the assumptions affects the constants in the conclusions. In particular, it can be seen from the analysis in Hairer et al. (2011) that growth of any of the parameters l, K, ϱ, ϑ causes an increase in the constant C and a decrease in the constant λ. In other words, enlarging any of the parameters l, K, ϱ, ϑ results in a worse spectral gap. This observation is relevant in order to obtain uniform spectral gaps for a sequence of Markov chains. Namely, suppose that we have Markov kernels $\{P_n\}_{n\in\mathbb{N}}$ (with perhaps different state spaces) for which we can find distance like functions $\{d_n\}_{n\in\mathbb{N}}$ and Lyupanov functionals $\{V_n\}_{n\in\mathbb{N}}$ satisfying the conditions in theorem 14 with constants $\tilde{l}, \tilde{K}, \tilde{\varrho}, \tilde{\vartheta}$ (independent of n). We can then deduce that the constants $\lambda > 0$ and $C > 0$ in the conclusion of the weak Harris theorem can be chosen independently of n. It is precisely this observation that is exploited in Hairer et al. (2014)

It is then important to highlight the main differences between our set-up and the one in Hairer et al. (2014). First, the Markov kernels that we consider in this paper are not defined on the same state space and in particular the log-likelihoods Φ_n, Φ, although related, are different. Secondly, our discretization of the continuum prior $\boldsymbol{\pi}$ is the prior $\boldsymbol{\pi_n}$ supported on $L^2(\gamma_n)$ and not the discretization constructed by truncating the Karhunen Loève expansion of the continuum prior. These differences in the set-ups, however, do not prevent us from using the proof of Theorem 4.7 in Hairer et al. (2014) thanks to the following three observations.

i) (Uniform control on local Lipschitz constants of log-likelihoods)

Lemma 16 *There exists a constant $L > 0$ such that for every $r > 0$ and $n \in \mathbb{N}$*

$$\sup_{u_n, v_n \in \mathcal{B}_r^n} \frac{|\Phi_n(u_n; y) - \Phi_n(v_n; y)|}{\|u_n - v_n\|} \leq Lr, \quad \sup_{u,v\in\mathcal{B}_r} \frac{|\Phi(u; y) - \Phi(v; y)|}{\|u - v\|} \leq Lr,$$

where in the above $\mathcal{B}_r^n$ ($\mathcal{B}_r$) denotes the ball in $L^2(\gamma_n)$ ($L^2(\gamma)$) centered at the origin and with radius r.

Proof Recall that

$$\Phi_n(u_n; y) = \phi^y(\mathcal{G}_n(u_n)), \quad u_n \in L^2(\gamma_n),$$

and so, thanks to Assumptions 1 on ϕ^y, we get

$$\begin{aligned}|\Phi_n(u_n; y) - \Phi_n(v_n; y)| &\leq |\phi^y(\mathcal{G}_n(u_n)) - \phi^y(\mathcal{G}_n(v_n))| \\ &\leq C_1 \max\{|\mathcal{G}_n(u_n)|, |\mathcal{G}_n(v_n)|, 1\}|\mathcal{G}_n(u_n) - \mathcal{G}_n(v_n)|.\end{aligned}$$

Now, recall that the vector $\mathcal{G}_n(u_n) - \mathcal{G}_n(v_n) \in \mathbb{R}^p$ has coordinates

$$[\mathcal{G}_n(u_n) - \mathcal{G}_n(v_n)]_i = \frac{1}{\gamma_n(B_\delta(\mathbf{x}_i))}\langle \mathbb{1}_{B_\delta(\mathbf{x}_i)}, \mathcal{F}_n(u_n) - \mathcal{F}_n(v_n)\rangle_{L^2(\gamma_n)}, \quad i = 1, \ldots, p.$$

From the Cauchy-Schwartz inequality it follows that

$$\begin{aligned} |[\mathcal{G}_n(u_n) - \mathcal{G}_n(v_n)]_i| &\leq \frac{1}{(\gamma_n(B_\delta(\mathbf{x}_i)))^{1/2}} \|\mathcal{F}_n(u_n) - \mathcal{F}_n(v_n)\|_{L^2(\gamma_n)} \\ &\leq \frac{1}{(\gamma_n(B_\delta(\mathbf{x}_i)))^{1/2}} \|u_n - v_n\|_{L^2(\gamma_n)}, \end{aligned}$$

where in the last line we have used the fact that $\mathcal{F}_n$ is a linear map as well as the fact that it is a contraction. Since

$$\gamma_n(B_\delta(\mathbf{x}_i)) \to \gamma(B_\delta(\mathbf{x}_i)), \qquad \text{as } n \to \infty, \tag{36}$$

it follows that

$$|\mathcal{G}_n(u_n) - \mathcal{G}_n(v_n)| \leq C_2 \|u_n - v_n\|_{L^2(\gamma)},$$

where C_2 is independent of $u_n, v_n \in L^2(\gamma_n)$ or $n \in \mathbb{N}$. Therefore, there exists a constant C_3 (independent of $u_n, v_n \in L^2(\gamma_n)$ or $n \in \mathbb{N}$) such that

$$|\Phi_n(u_n; y) - \Phi_n(v_n; y)| \leq C_3 \max\{\|u_n\|_{L^2(\gamma_n)}, \|v_n\|_{L^2(\gamma_n)}, 1\} \|u_n - v_n\|_{L^2(\gamma_n)}.$$

Naturally the same analysis holds for Φ and this finishes the proof. ■

Remark 17 *The conclusions in the previous lemma hold for non-linear forward maps $\mathcal{F}_n$, $\mathcal{F}$ that are (uniformly in n) Lipschitz and have (uniformly in n) linear growth.*

ii) (Dominating limiting measure) We make use of a "limiting measure" that dominates the measures $\boldsymbol{\pi_n}$ in the sense described below. Notice that we cannot use the continuum prior $\boldsymbol{\pi}$, but a slight modification of it will suffice.

Lemma 18 *There exists a large enough $\rho > 0$, such that the Gaussian measure*

$$\boldsymbol{\pi}^\rho := N\big(0, (1+\rho)^2 (\alpha I - \Delta_{\mathcal{M}})^{-s}\big),$$

satisfies

$$\int_{L^2(\gamma_n)} g(\|u_n\|_{L^2(\gamma_n)}) d\boldsymbol{\pi_n}(u_n) \leq \int_{L^2(\gamma)} g(\|u\|_{L^2(\gamma)}) d\boldsymbol{\pi}^\rho(u),$$

for every $n \in \mathbb{N}$ and every increasing function $g : [0, \infty) \to \mathbb{R}$. In particular, for every $r > 0$ and every $n \in \mathbb{N}$,

$$\boldsymbol{\pi_n}\left(L^2(\gamma_n) \setminus \mathcal{B}_r^n\right) \leq \boldsymbol{\pi}^\rho\left(L^2(\gamma) \setminus \mathcal{B}_r\right).$$

Proof Thanks to inequality (7), we can find $\rho > 0$ such that, for every $n \in \mathbb{N}$,

$$\frac{1}{(\alpha + \lambda_i^n)^s} \leq \frac{1+\rho}{(\alpha + \lambda_i)^s}, \quad \forall i = 1, \dots, k_n.$$

Using the Karhunen Loève expansion to represent random variables with laws $\boldsymbol{\pi_n}$ and $\boldsymbol{\pi}^\rho$ we can easily deduce the inequality for the measures of complements of balls (last inequality). The inequality for a general increasing function g follows from a standard approximation with increasing step functions. ■

iii) (Uniform lower bound for acceptance probability) The next lemma provides uniform control on the acceptance probability of the pCN algorithm when a proposal lies within a fixed distance of a contracted version of the current state of the chain. More precisely:

Lemma 19 *Let $a(u,v)$ be the acceptance probability in Algorithm 1 for continuum pCN and $a_n(u_n,v_n)$ the acceptance probability in Algorithm 2 for graph pCN. Fix an arbitrary $r>0$. Then, there exists $c\in\mathbb{R}$ such that*

$$\inf_{w_n\in\mathcal{B}_r^n(\sqrt{1-\beta^2}v_n)} a_n(v_n,w_n)\geq \exp(c)>0, \qquad \inf_{w\in\mathcal{B}_r(\sqrt{1-\beta^2}v)} a(v,w)\geq \exp(c)>0$$

for arbitrary $v_n\in L^2(\gamma_n)$, $v\in L^2(\gamma)$ and $n\in\mathbb{N}$.

Proof First of all notice that

$$\|\mathcal{G}_n\|\leq\|\mathcal{O}_n\|\|\mathcal{F}_n\|\leq\|\mathcal{O}_n\|,$$

where in the last inequality we have used that $\mathcal{F}_n$ is a contraction. Thanks to (36) it follows that

$$\|\mathcal{O}_n\|\to\|\mathcal{O}\|, \quad \text{as } n\to\infty,$$

and in particular we can find a constant $\tilde{K}$ (independent of n) such that

$$\|\mathcal{G}_n\|\leq\tilde{K}.$$

Let $v_n,w_n\in L^2(\gamma_n)$ be such that $w_n\in\mathcal{B}_r^n(\sqrt{1-\beta^2}v_n)$. Then,

$$\begin{aligned}|\mathcal{G}_n(w_n)-\sqrt{1-\beta^2}\mathcal{G}_n(v_n)| &= |\mathcal{G}_n(w_n-\sqrt{1-\beta^2}v_n)|\\ &\leq\|\mathcal{G}_n\|\|w_n-\sqrt{1-\beta^2}v_n\|_{L^2(\gamma_n)}\\ &\leq\tilde{K}r=:K.\end{aligned}$$

From Assumptions 1 we deduce that

$$\Phi_n(v_n;y)-\Phi_n(w_n;y)=\phi^y(\mathcal{G}_n(v_n))-\phi^y(\mathcal{G}_n(w_n))\geq c,$$

for a c that is independent of n. Hence,

$$\inf_{w_n\in\mathcal{B}_r^n(\sqrt{1-\beta^2}v_n)} a(v_n,w_n)\geq\exp(c)>0.$$

Naturally the same analysis holds for Φ and this finishes the proof. ■

Remark 20 *The same conclusions in the previous lemma hold for non-linear forward maps $\mathcal{F}_n$, $\mathcal{F}$ that are (uniformly in n) Lipschitz, have (uniformly in n) linear growth, and are positively homogeneous of degree one.*

Proof [Proof of Theorem 7] Lemmas 16, 18 and 19 allow us to follow the analysis in Hairer et al. (2014) (where in our case we use $\boldsymbol{\pi}^{\rho}$ from Lemma 18) and check that the conditions of the weak Harris theorem (with distance like functional d and Lyapunov functional V as in the statement of our theorem) are satisfied with constants l, K, ϱ, ϑ that are independent of the discretization.

■

Proof [Proof of Corollary 10] By Proposition 2.8 and Lemma 2.9 in Hairer et al. (2014), and the reversibility of the Markov kernel of the pCN algorithm, it is enough to check that the space

$$Lip(\tilde{d}) \cap L^{\infty}(\mathcal{H};\mu),$$

is dense in $L^2(\mathcal{H};\mu)$. Here $\tilde{d}$ denotes the distance-like function from Theorem 7 and μ stands for the invariant measure of the Markov chain (in this case the posterior distribution). In the finite dimensional case (i.e. $\mathcal{H} = L^2(\gamma_n)$) this is a simple consequence of a standard mollification argument. More precisely, it follows from the following observations:

i) For every $R > 0$, $\|\cdot\|$-Lispchitz functions on $\mathcal{B}^n_R$ are also $\tilde{d}$-Lipschitz on $\mathcal{B}^n_R$.

ii) $\|\cdot\|$-Lispchitz functions on $\mathcal{B}^n_R$ are dense in $L^2(\mathcal{B}^n_R;\mu)$ (by mollification).

iii) $f \in L^2(\mathcal{H};\mu)$ can be approximated with $\{f_k\}_{k\in\mathbb{N}}$, where

$$f_k(u) := \eta_k(\|u\|)\min\{\max\{f(u), -k\}, k\}, \quad u \in \mathcal{H},$$

where $\eta_k : [0,\infty) \to [0,1]$ is a smooth cut-off function which satisfies $\eta_k(r) = 1$ if $r < k$ and $\eta_k(r) = 0$ if $r > 2k$.

For the infinite dimensional case it is enough to reduce the problem to the finite dimensional case. This reduction is achieved as follows. Without the loss of generality an arbitrary element $u \in \mathcal{H}$ can be written as $u = (u_1, u_2, \dots)$ and for every $k \in \mathbb{N}$ we may consider the projection:

$$\Pi^c_k : u \in \mathcal{H} \mapsto (u_{k+1}, u_{k+2}, \dots),$$

and the measure $\mu^c_k := \Pi^c_{k\sharp}\mu$. For an arbitrary $f \in L^2(\mathcal{H};\mu)$, we can then define the sequence $\{f_k\}_{k\in\mathbb{N}} \subseteq L^2(\mathcal{H};\mu)$ defined by

$$f_k(u) := \int f(u_1, \dots, u_k, v_{k+1}, v_{k+2}, \dots) d\mu^c_k(v_{k+1}, v_{k+2}, \dots), \quad u \in \mathcal{H}$$

Notice that for each k the function f_k depends only on the first k coordinates of u and so we can apply the result for the finite dimensional case to approximate f_k with functions in $Lip(\tilde{d}) \cap L^{\infty}(\mathcal{H};\mu)$. From the straightforward fact that $f_k \to_{L^2(\mathcal{H};\mu)} f$, the approximation of functions in $L^2(\mathcal{H};\mu)$ with functions in $Lip(\tilde{d}) \cap L^{\infty}(\mathcal{H};\mu)$ now follows.

■

Appendix E. Verification of Hypotheses for Gaussian and Probit Noise Models

E.1. Gaussian

Let us show that the Gaussian model satisfies Assumption 1.

i) Let $K > 0$ and let $\tau > 0$ be such that $(1-\tau) > (1+\tau)(1-\beta^2)$. For such τ choose $R = R_\tau > 0$ large enough so that if $u \in \mathbb{R}^p$ satisfies $|u| \geq R$ then

$$(1-\tau)|u|^2 \leq |u-y|^2 \leq (1+\tau)|u|^2.$$

Let $v, w \in \mathbb{R}^p$ be such that $|w - \sqrt{1-\beta^2}v| \leq K$. If $|w| \leq R + K$, then

$$|v-y|^2 - |w-y|^2 \geq 0 - 2|y|^2 - 2(R+K)^2.$$

On the other hand, if $|w| \leq R + K$, we see that

$$R + K \leq |w| \leq \sqrt{1-\beta^2}|v| + K,$$

and it follows that

$$\begin{aligned} |v-y|^2 - |w-y|^2 &\geq (1-\tau)|v|^2 - (1+\tau)|w|^2 \\ &\geq ((1-\tau) - (1+\tau)(1-\beta^2))|v|^2 - 2(1+\tau)\sqrt{1-\beta^2}K|v| - (1+\tau)K^2 \\ &\geq C_1, \end{aligned}$$

for some real number C_1.

From the above analysis we deduce that for every v, w with $|w - \sqrt{1-\beta^2}\, v| \leq K$,

$$\phi^y(v) - \phi^y(w) \geq c,$$

for some $c \in \mathbb{R}$.

ii) The second assumption is easily seen to be satisfied by the Gaussian model.

E.2. Probit

Let us show that the probit model satisfies Assumption 1.

i) Let $K > 0$ and consider $v, w \in \mathbb{R}^p$ such that $|w - \sqrt{1-\beta^2}v| \leq K$. Then,

$$|y_i w_i - \sqrt{1-\beta^2} y_i v_i| = |w_i - \sqrt{1-\beta^2} v_i| \leq |v - \sqrt{1-\beta^2}\, w| \leq K, \quad i = 1, \dots, p, \tag{37}$$

where the first equality follows from the fact that $y_i \in \{-1, 1\}$. In particular,

$$|y_i w_i| \leq K + \sqrt{1-\beta^2}|y_i v_i|, \quad i = 1, \dots, p. \tag{38}$$

Notice that the function $t \in \mathbb{R} \mapsto -\log(\Psi(t))$ is decreasing. Hence, if $y_i w_i > -(1/(1-\sqrt{1-\beta^2}) + 1)K$ we see that

$$-\log(\Psi(y_i v_i)) - (-\log(\Psi(y_i v_i))) \geq 0 + \log(\Psi(-(1/(1-\sqrt{1-\beta^2}) + 1)K)).$$

On the other hand, if $y_i w_i < -(1/(1-\sqrt{1-\beta^2})+1)K$ we deduce from (37) that $y_i v_i < -K/(1-\sqrt{1-\beta^2}) < 0$ and from (38) we deduce

$$y_i v_i \leq \sqrt{1-\beta^2} y_i v_i - K \leq y_i w_i,$$

from where it follows that

$$-\log(\Psi(y_i v_i)) - (-\log(\Psi(y_i v_i))) \geq 0.$$

From the above analysis we deduce that, for every v, w with $|w - \sqrt{1-\beta^2} v| \leq K$,

$$\phi^y(v) - \phi^y(w) \geq c := p \log(\Psi(-(1/(1-\sqrt{1-\beta^2})+1)K)).$$

ii) Let us now check that the probit model satisfies the second assumption on ϕ^y. Since the function

$$g : t \in \mathbb{R} \mapsto -\log(\Psi(t))$$

is decreasing, convex, and converges to zero as $t \to \infty$, the first assumption on ϕ^y will hold if we can show that

$$\limsup_{t \to -\infty} \frac{|g'(t)|}{|t|} < \infty.$$

This however follows from the fact that

$$g'(t) = \frac{-e^{-t^2/2}}{\int_{-\infty}^{t} e^{-r^2/2} dr},$$

and the well known fact that

$$\frac{e^{-t^2/2}}{2|t|} \leq \int_{-\infty}^{t} e^{-r^2/2} dr,$$

for all negative enough t.