

UNCERTAINTY QUANTIFICATION IN THE CLASSIFICATION OF HIGH DIMENSIONAL DATA *

ANDREA L BERTOZZI [†], XIYANG LUO [‡], ANDREW M. STUART [§], AND KONSTANTINOS C. ZYGALAKIS [¶]

Abstract. Classification of high dimensional data finds wide-ranging applications. In many of these applications equipping the resulting classification with a measure of uncertainty may be as important as the classification itself. In this paper we introduce, develop algorithms for, and investigate the properties of, a variety of Bayesian models for the task of binary classification; via the posterior distribution on the classification labels, these methods automatically give measures of uncertainty. The methods are all based around the graph formulation of semi-supervised learning.

We provide a unified framework which brings together a variety of methods which have been introduced in different communities within the mathematical sciences. We study probit classification [43], generalize the level-set method for Bayesian inverse problems [24] to the classification setting, and generalize the Ginzburg-Landau optimization-based classifier [5, 40] to a Bayesian setting; we also show that the probit and level set approaches are natural relaxations of the harmonic function approach introduced in [49]. We introduce efficient numerical methods, suited to large data-sets, for both MCMC-based sampling as well as gradient-based MAP estimation. Through numerical experiments we study classification accuracy and uncertainty quantification for our models; these experiments showcase a suite of datasets commonly used to evaluate graph-based semi-supervised learning algorithms.

Key words. Graph classification, Uncertainty quantification, Gaussian prior

AMS subject classifications. 6209, 65S05, 9404

1. Introduction.

1.1. The Central Idea. Semi-supervised learning has attracted the attention of many researchers because of the importance of combining unlabeled data with labeled data. In many applications the number of unlabeled data points is so large that labeling training data is expensive and time-consuming. Therefore, the problem of effectively utilizing a combination of unlabeled and labeled information is very important in machine learning research. This paper concerns the issue of how to address uncertainty quantification in such classification methods. In doing so we bring together a variety of themes from the mathematical sciences, including optimization, PDEs, probability and statistics. We will show that a variety of different methods, arising in very distinct communities, can all be formulated around a common objective function

$$J(w) = \frac{1}{2} \langle w, Pw \rangle + \Phi(w)$$

*Submitted to the editors DATE.

Funding: NSF grant DMS-1118971, NSF grant DMS-1417674, and ONR grant N00014-16-1-2119. AMS is funded by DARPA and EPSRC. KCZ was partially supported by a grant from the Simons Foundation and by the Alan Turing Institute under the EPSRC grant EP/N510129/1. Part of this work was done during the author's stay at the Newton Institute for the program Stochastic Dynamical Systems in Biology: Numerical Methods and Applications.

[†]Department of Mathematics, University of California Los Angeles, Los Angeles, CA (bertozzi@math.ucla.edu).

[‡]Department of Mathematics, University of California Los Angeles, Los Angeles, CA (math-luo@math.ucla.edu).

[§]Computing and Mathematical Sciences, Caltech, Pasadena, CA (astuart@caltech.edu).

[¶]School of Mathematics, University of Edinburgh, Edinburgh, Scotland (k.zygalakis@ed.ac.uk).

for a real valued function w on the nodes of a graph representing the data points. The matrix P is proportional to a graph Laplacian derived from the unlabeled data and the function Φ involves the labelled data. The variable w is used for classification. Minimizing this objective function is one approach to such a classification. A probability distribution related to the objective function has density $\mathbb{P}(w)$ proportional to $\exp(-J(w))$; probability of the labelling variable w is high where the objective function is small, and vice-versa. Uncertainty quantification corresponds to using the probability distribution to compute expectations of test functions g , defined on the nodes of the graph, which enable us to measure the variability of label variables, such as means and variances:

$$\int g(w)\mathbb{P}(w)dw.$$

In the settings of interest this will typically be a very high dimensional integral, with the dimension given by the number of unlabelled data points. Carrying out this program requires computational algorithms to minimize $J(w)$ or to draw samples, via Monte Carlo Markov chain (MCMC) for example, from the probability distribution with density $\mathbb{P}(w)$. These algorithms exploit the fact that $\frac{1}{2}\langle w, Pw \rangle$ is a graph analogue of the Dirichlet energy and will leverage analogies with PDE-based methodologies involving the classical Euclidean Dirichlet energy in order to derive effective computational methods. In this paper we will describe this confluence of ideas from different parts of the mathematical sciences, show how our approach builds on a broad range of advances in the field which we will review, and demonstrate the emergence of a problem area with many open challenges for the mathematical sciences. We emphasize that the variety of probabilistic models considered in this paper arise from different assumptions concerning the structure of the data. Our objective is not to assess the validity of these assumptions, which is a modelling question best addressed on a case-by-case basis, but rather we develop an overarching computational framework suitable for all the models arising from these different assumptions.

1.2. Literature Review. An effective method for semi-supervised learning is to construct a similarity graph on both the unlabeled and labeled examples, and classify unknown labels by leveraging the graph structure. A central conceptual issue in the setting of this problem is that labels are discrete, whilst similarity information is often continuous. Strategies to work with both of these settings simultaneously are at the heart of this subject. In [8], Blum et al. posed the binary semi-supervised classification problem using a graph min-cut problem. This is equivalent to a maximum a posteriori (MAP) estimator with respect to a Bayesian posterior distribution for a Markov random field (MRF) over the discrete state space of binary labels [48]; the resulting optimization problem can be solved exactly in polynomial time. In general, inference for multi-label discrete MRFs is intractable [16]. However, several approximate algorithms exist for the multi-label case [10, 9, 28], and have been applied to many imaging tasks [11, 4, 27].

The probit classification method, using Gaussian process priors, is described in [43]; however in that book the prior does not depend on the unlabelled data. Gaussian priors which depend on the unlabelled data may be constructed by using the Graph Laplacian, an approach undertaken in [25, 20, 47, 49, 50]. The model defined in [49] is equivalent to a continuum relaxation of the discrete state space MRF in [8]. The Bayesian formulation which underpins our work in this paper was made explicit in [25, 50] where a variety of likelihood models are used to condition on the labelled data; the probit approach, for example, could be used to accomplish this. Probit

utilizes the same prior as in [49] but the data is assumed to take binary values, found from thresholding the underlying continuous variable, and thereby providing a link between the combinatorial and continuous state space approaches described in the previous paragraph. The probit methodology is often implemented via MAP optimization – that is the posterior probability is maximized rather than sampled – or an approximation to the posterior is computed, in the neighbourhood of the MAP estimator [43]. For full posterior exploration, Gibbs sampling is often used [1] and this methodology has been applied recently in [20]; furthermore methods designed to break undesirable dependencies in the Gibbs sampler are introduced in [22]. In the context of MAP estimation, the graph-based terms act as a regularizer, in the form of the graph Dirichlet energy $\frac{1}{2}\langle w, Pw \rangle$. A formal framework for graph-based regularization can be found in [2, 3]. More recently, other forms of regularization have been considered such as the graph wavelet regularization [36, 19].

Another link between discrete combinatorial optimization approaches and methods based on optimization over real-valued variables was made in the work of Bertozzi et al. [5, 40]. The approach is based on the fact that the TV functional, when suitably generalized to weighted graphs, coincides with the graph cut energy. Relaxation of the TV functional is well-understood in the context of partial differential equations (PDE) and generalizing ideas applicable to the PDE Laplacian in the context of the graph Laplacian leads to new optimization methods. Based on this reasoning, in [5] the graph Ginzburg-Landau functional was used as a relaxation of the graph TV functional for the task of binary classification. This was generalized to multi-class classification in [18]. Following this line of work, several new algorithms were developed for semi-supervised and unsupervised classification problems on weighted graphs [23, 29]. A further connection with PDE based methods is the level-set approach to Bayesian inversion, introduced recently in [24]; this is very closely related to our variant on the probit method, as we will demonstrate.

There are a wide range of methodologies employed in the field of uncertainty quantification, and the reader may consult the books [37, 38, 45] and the recent article [33] for details and further references. Underlying all of these methods is a Bayesian methodology which is attractive both for its clarity with respect to modelling assumptions and its basis for application of a range of computational tools. Nonetheless it is important to be aware of limitations in this approach, in particular with regard to its robustness with respect to the specification of the model, and in particular the prior distribution on the unknown of interest [32]. Whilst the book [43] conducts a number of thorough uncertainty quantification studies for a variety of learning problems using Gaussian process priors, most of the papers studying graph based learning referred to above primarily use the Bayesian approach to learn hyperparameters in an optimization context, and do not consider uncertainty quantification.

1.3. Our Contribution. In this paper, we focus exclusively on the problem of binary semi-supervised classification; however the methodology and conclusions will extend beyond this setting. Our focus is on a presentation which puts uncertainty quantification at the heart of the problem formulation, and we make four primary contributions:

- we define a number of different Bayesian formulations of the graph-based semi-supervised learning problem and we connect them to one another, to binary classification methods and to a variety of PDE-inspired approaches to classification; in so doing we provide a single framework for a variety of methods which have arisen in distinct communities and we open up a number

of new avenues of study for the problem area;

- we highlight the pCN-MCMC method for posterior sampling which, based on analogies with its use for PDE-based inverse problems [15], has the potential to sample the posterior distribution in a number of steps which is independent of the number of graph nodes;
- we introduce approximations exploiting the empirical properties of the spectrum of the graph Laplacian, generalizing methods used in the optimization context in [5], allowing for computations at each MCMC step which scale well with respect to the number of graph nodes;
- we demonstrate, by means of numerical experiments on a range of problems, both the feasibility, and value, of Bayesian uncertainty quantification in semi-supervised, graph-based, learning.

1.4. Overview and Notation. The paper is organized as follows. In section 2, we give some background material needed for problem specification. In section 3 we formulate the four Bayesian models used for the classification tasks. Section 4 introduces the MCMC and optimization algorithms that we use. In section 5, we present and discuss results of numerical experiments to illustrate our findings; these are based on four examples of increasing size: the house voting records from 1984 (as used in [5]), the tuneable two moons data set [13], the MNIST digit data base [26] and the hyperspectral gas plume imaging problem [12]. We conclude in section 6.

To aid the reader, we give here an overview of notation used throughout the paper.

- Z the set of nodes of the graph, with cardinality N ;
- Z' the set of nodes where labels are observed, with cardinality $J \leq N$;
- $x : Z \mapsto \mathbb{R}^d$, feature vectors;
- $u : Z \mapsto \mathbb{R}$ latent variable characterizing nodes, with $u(j)$ denoting evaluation of u at node j ;
- $S : \mathbb{R} \mapsto \{-1, 1\}$ the thresholding function;
- S_ϵ relaxation of S using gradient flow in double-well potential W_ϵ ;
- $l : Z \mapsto \{-1, 1\}$ the label value at each node with $l(j) = S(u(j))$;
- $y : Z' \mapsto \{-1, 1\}$ or $y : Z' \mapsto \mathbb{R}$, label data;
- $v : Z \mapsto \mathbb{R}$ with v being a relaxation of the label variable l ;
- A weight matrix of the graph, L the resulting symmetric graph Laplacian;
- P the precision matrix and C the covariance matrix, both found from L ;
- $\{q_k, \lambda_k\}_{k=0}^{N-1}$ eigenpairs of L ;
- U : orthogonal complement of the null space of the graph Laplacian L , given by q_0^\perp ;
- U_ℓ : orthogonal complement of the first ℓ eigenfunctions of the graph Laplacian L .
- $|\cdot|$ denotes the Euclidean norm and $\langle \cdot, \cdot \rangle$ the corresponding inner-product;
- GL : Ginzburg-Landau functional;
- μ_0, ν_0 : prior probability measures;
- μ and ν : (with suffices denoting different models);
- the measures denoted μ typically take argument u and are real-valued; the measures denoted ν take argument l on label space, or argument v on a real-valued relaxation of label space;
- $\mathcal{N}(m, \Sigma)$ denotes a Gaussian random variable with mean m and covariance Σ ;
- \mathbb{P} and \mathbb{E} denote the probability of an event, and the expectation of a ran-

dom variable, respectively; the underlying probability measure will be made explicit as a subscript when it is necessary to do so.

2. Problem Specification. In subsection 2.1 we formulate semi-supervised learning as a problem on a graph. Subsection 2.2 defines the relevant properties of the graph Laplacian and in subsection 2.3 these properties are used to construct a Gaussian probability distribution; in section 3 this Gaussian will be used to define our prior information about the classification problem. In subsection 2.4 we discuss thresholding which provides a link between the real-valued prior information, and the label data provided for the semi-supervised learning task; in section 3 this will be used to define our likelihood.

2.1. Semi-Supervised Learning on a Graph. We are given a set of points denoted by $Z = \{1, \dots, N\}$, and a set of features $X = \{x_1, \dots, x_N\}$ associated with these points; each feature vector x_j is an element of \mathbb{R}^d , so that $X \in \mathbb{R}^{d \times N}$. Graph learning starts from the construction of an undirected graph G with weights a_{ij} computed from the feature set X . For graph semi-supervised learning, we are also given a partial set of (possibly noisy) labels $y = \{y(j) | j \in Z'\}$, where $Z' \subseteq Z$ has size $J \leq N$. The task is to infer the labels for all nodes in Z , using the weighted graph G and also the set of noisily observed labels y . In the Bayesian formulation which we adopt the feature set X , and hence the graph G , is viewed as prior information, describing correlations amongst the nodes of the graph, and we combine this with a likelihood based on the noisily observed labels y , to obtain a posterior distribution on the labelling of all nodes. Various Bayesian formulations, which differ in the specification of the observation model and/or the prior, are described in section 3. In the remainder of this section we give the background needed to understand all of these formulations, thereby touching on the graph Laplacian itself, its link to Gaussian probability distributions and, via thresholding, to non-Gaussian probability distributions and to the Ginzburg-Landau functional. An important point to appreciate is that building our priors from Gaussians confers considerable computational advantages for large graphs; for this reason the non-Gaussian priors will be built from Gaussians via change of measure or push forward under a nonlinear map.

2.2. The Graph Laplacian. The graph Laplacian is central to many graph-learning algorithms. There are a number of variants used in the literature; see [5, 41] for a discussion. We will work with the symmetric Laplacian, defined from the weight matrix $A = \{a_{ij}\}$ as follows. We define the diagonal matrix $D = \text{diag}\{d_{ii}\}$ with entries $d_{ii} = \sum_{j \in Z} a_{ij}$. If we assume that the graph G is connected, then $d_{ii} > 0$ for all nodes $i \in Z$. We can then define the symmetric graph Laplacian¹ as

$$(1) \quad L = I - D^{-1/2} A D^{-1/2},$$

and the graph Dirichlet energy as $J_0(u) := \frac{1}{2} \langle u, Lu \rangle$. Then

$$(2) \quad J_0(D^{\frac{1}{2}} u) = \frac{1}{4} \sum_{\{i,j\} \in Z \times Z} a_{ij} (u(i) - u(j))^2.$$

¹In the majority of the paper the only property of L that we use is that it is symmetric positive semi-definite. We could therefore use other graph Laplacians, such as the unnormalized choice $L = D - A$, in most of the paper. The only exception is the spectral approximation sampling algorithm introduced later; that particular algorithm exploits empirical properties of the symmetrized graph Laplacian. Note, though, that the choice of which graph Laplacian to use can make a significant difference – see [5], and Figure 2.1 therein. To make our exposition more concise we confine our presentation to the symmetric graph Laplacian.

Thus, similarly to the classical Dirichlet energy, this quadratic form penalizes nodes from having different function values, with penalty being weighted with respect to the similarity weights from A . Furthermore the identity shows that L is positive semi-definite. Indeed the vector of ones $\mathbb{1}$ is in the null-space of $D - A$ by construction, and hence L has a zero eigenvalue with corresponding eigenvector $D^{\frac{1}{2}}\mathbb{1}$.

We let (q_k, λ_k) denote the eigenpairs of the matrix L , ordered so that

$$\lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{N-1} \leq 2.$$

The upper bound of 2 may be found in [14, Lemma 1.7, Chapter 1]. The eigenvector corresponding to $\lambda_0 = 0$ is $q_0 = D^{\frac{1}{2}}\mathbb{1}$ and $\lambda_1 > 0$, assuming a fully connected graph. Then $L = Q\Lambda Q^*$ where Q has columns $\{q_k\}_{k=0}^{N-1}$ and Λ is a diagonal matrix with entries $\{\lambda_k\}_{k=0}^{N-1}$. Using these eigenpairs the graph Dirichlet energy can be written as

$$(3) \quad \frac{1}{2}\langle u, Lu \rangle = \frac{1}{2} \sum_{j=1}^{N-1} \lambda_j (\langle u, q_j \rangle)^2;$$

this is analogous to decomposing the classical Dirichlet energy using Fourier analysis.

2.3. Gaussian Measure. We now show how to build a Gaussian distribution with negative log density proportional to $J_0(u)$. Such a prior prefers functions that have larger components on the first few eigenvectors of the graph Laplacian, where the eigenvalues of L are smaller. The corresponding eigenvectors carry rich geometric information about the weighted graph. For example, the second eigenvector of L is the *Fiedler vector* and solves a relaxed normalized min-cut problem [41, 21]. The Gaussian distribution thereby connects geometric intuition embedded within the graph Laplacian to a natural probabilistic picture.

To make this connection concrete we define diagonal matrix Σ with entries defined by the vector

$$(0, \lambda_1^{-1}, \dots, \lambda_{N-1}^{-1})$$

and define the positive semi-definite covariance matrix $C = cQ\Sigma Q^*$; choice of the scaling c will be discussed below. We let $\mu_0 := \mathcal{N}(0, C)$. Note that the covariance matrix is that of a Gaussian with variance proportional to λ_j^{-1} in direction q_j thereby leading to structures which are more likely to favour the Fiedler vector ($j = 1$), and lower values of j in general, than it does higher values. The fact that the first eigenvalue of C is zero ensures that any draw from μ_0 changes sign, because it will be orthogonal to q_0 .² To make this intuition explicit we recall the Karhunen-Loeve expansion which constructs a sample u from the prior μ_0 according to the random sum

$$(4) \quad u = c^{\frac{1}{2}} \sum_{j=1}^{N-1} \lambda_j^{-\frac{1}{2}} q_j z_j,$$

where the $\{z_j\}$ are i.i.d. $\mathcal{N}(0, 1)$. Since each q_j with $j \geq 1$ is orthogonal to q_0 it follows that u is orthogonal to q_0 and the sign-change property is enforced because q_0 is of one sign.

²Other choices of the first eigenvalue are possible and may be useful but for simplicity of exposition we do not consider them in this paper.

We choose the constant of proportionality c as a rescaling which enforces the property $\mathbb{E}|u|^2 = N$ for $u \sim \mu_0 := \mathcal{N}(0, C)$; in words the per-node variance is 1. Note that, using the orthogonality of the $\{q_j\}$,

$$\mathbb{E}|u|^2 = c \sum_{j=1}^{N-1} \lambda_j^{-1} \mathbb{E}z_j^2 = c \sum_{j=1}^{N-1} \lambda_j^{-1}.$$

Thus the normalization implies that

$$(5) \quad c = N \left(\sum_{j=1}^{N-1} \lambda_j^{-1} \right)^{-1}.$$

We reiterate that the support of the measure μ_0 is the space $U := q_0^\perp = \text{span}\{q_1, \dots, q_{N-1}\}$ and that, on this space, the probability density function is proportional to

$$\exp\left(-c^{-1}J_0(u)\right) = \exp\left(-\frac{1}{2c}\langle u, Lu \rangle\right),$$

so that the *precision matrix* of the Gaussian is $P = c^{-1}L$. In what follows the sign of u will be related to the classification; since all the entries of q_0 are positive, working on the space U ensures a sign change in u , and hence a non-trivial classification.

2.4. Thresholding and Non-Gaussian Probability Measure. For the models considered in this paper, the label space of the problem is discrete while the latent variable u through which we will capture the correlations amongst nodes of the graph, encoded in the feature vectors, is real-valued. We describe thresholding, and a relaxation of thresholding, to address the need to connect these two differing sources of information about the problem. In what follows the latent variable $u : Z \rightarrow \mathbb{R}$ is thresholded to obtain the label variable $l : Z \rightarrow \{-1, 1\}$. The variable $v : Z \rightarrow \mathbb{R}$ is a real-valued relaxation of the label variable l . The variable u will be endowed with a Gaussian probability distribution. From this the variable l (which lives on a discrete space) and v (which is real-valued, but concentrates near the discrete space supporting l) will be endowed with non-Gaussian probability distributions.

Define the (signum) function $S : \mathbb{R} \mapsto \{-1, 1\}$ by

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

This will be used to connect the latent variable u with the label variable l . The function S may be relaxed by defining $S_\epsilon(u) = v|_{t=1}$ where v solves the gradient flow

$$\dot{v} = -\nabla W_\epsilon(v), \quad v|_{t=0} = u \quad \text{for potential} \quad W_\epsilon(v) = \frac{1}{4\epsilon}(v^2 - 1)^2.$$

This will be used, indirectly, to connect the latent variable u with the real-valued relaxation of the label variable, v . Note that $S_\epsilon(\cdot) \rightarrow S(\cdot)$, pointwise, as $\epsilon \rightarrow 0$, on $\mathbb{R} \setminus \{0\}$. This reflects the fact that the gradient flow minimizes W_ϵ , asymptotically as $t \rightarrow \infty$, whenever started on $\mathbb{R} \setminus \{0\}$.

We have introduced a Gaussian measure μ_0 on the latent variable u which lies in $U \subset \mathbb{R}^N$; we now want to introduce two ways of constructing non-Gaussian measures on the label space $\{-1, 1\}^N$, or on real-valued relaxations of label space, building on the measure μ_0 . The first is to consider the push-forward of measure μ_0 under the map S : $S^\# \mu_0$. Then

$$(S^\# \mu_0)(l) = \mu_0\left(S(u(j)) = l(j), 1 \leq j \leq |Z|\right).$$

Thus $S^\sharp \mu_0$ is a measure on the label space $\{-1, 1\}^N$. The second approach is to work with a change of measure from the Gaussian μ_0 in such a way that the probability mass on $U \subset \mathbb{R}^N$ concentrates close to the label space $\{-1, 1\}^N$. We may achieve this by defining the measure ν_0 via its Radon-Nykodim derivative

$$(6) \quad \frac{d\nu_0}{d\mu_0}(v) \propto e^{-\sum_{j \in Z} W_\epsilon(v(j))}.$$

We name ν_0 the Ginzburg-Landau measure, since the negative log density function of ν_0 is the graph Ginzburg-Landau functional

$$(7) \quad \text{GL}(v) := \frac{1}{2c} \langle v, Lv \rangle + \sum_{j \in Z} W_\epsilon(v(j)).$$

The Ginzburg-Landau distribution defined by ν_0 can be interpreted as a non-convex ground relaxation of the discrete MRF model [48], in contrast to the convex relaxation which is the Gaussian Field [49]. Since the double well has minima at the label values $\{-1, 1\}$, the probability mass of ν_0 is concentrated near the modes ± 1 , and ϵ controls this concentration effect.

3. Bayesian Formulation. In this section we formulate four different Bayesian models for the semi-supervised learning problem. The four models all combine the ideas described in the previous section to define four distinct posterior distributions. It is important to realize that these different models will give different answers to the same questions about uncertainty quantification, just as different methods based around optimization will give different classifications. The choice of which Bayesian model to use is related to the data itself, and making this choice is beyond the scope of this paper. Currently the choice must be addressed on a case by case basis, as is done when choosing an optimization method for classification. Nonetheless we will demonstrate that the shared structure of the four models mean that a common algorithmic framework can be adopted and we will make some conclusions about the relative costs of applying this framework to the four models.

We denote the latent variable by $u(j)$, $j \in Z$, the thresholded value of $u(j)$ by $l(j) = S(u(j))$ which is interpreted as the label assignment at each node j , and noisy observations of the binary labels by $y(j)$, $j \in Z'$. The variable $v(j)$ will be used to denote the real-valued relaxation of $l(j)$ used for the Ginzburg-Landau model. Recall Bayes formula which transforms a prior density $\mathbb{P}(u)$ on a random variable u into a posterior density $\mathbb{P}(u|y)$ on the conditional random variable $u|y$:

$$\mathbb{P}(u|y) = \frac{1}{\mathbb{P}(y)} \mathbb{P}(y|u) \mathbb{P}(u).$$

We will now apply this formula to condition our graph latent variable u , whose thresholded values correspond to labels, on the noisy label data y given at Z' . As prior on u we will always use $\mathbb{P}(u)du = \mu_0(du)$; we will describe three different likelihoods. We will also apply the formula to condition relaxed label variable v , on the same label data y , via the formula

$$\mathbb{P}(v|y) = \frac{1}{\mathbb{P}(y)} \mathbb{P}(y|u) \mathbb{P}(v).$$

We will use as prior the non-Gaussian $\mathbb{P}(v)dv = \nu_0(dv)$.

For the probit, level-set and atomic models, we now explicitly state the prior density $\mathbb{P}(u)$, the likelihood function $\mathbb{P}(y|u)$, and the posterior density $\mathbb{P}(u|y)$; in the

Ginzburg-Landau case v will replace u and we will define the densities $\mathbb{P}(v), \mathbb{P}(y|v)$ and $\mathbb{P}(v|y)$. Prior and posterior probability measures associated with letter μ are on the latent variable u ; measures associated with letter ν are on the label space, or real-valued relaxation of the label space.

3.1. Probit. The probit method is designed for classification and is described in [43]; in that context Gaussian process priors are used and these do not depend on label data. A recent fully Bayesian treatment of the methodology using unweighted graph Laplacians may be found in the paper [20]. In detail our model is as follows.

Prior We take as prior on u the Gaussian μ_0 . Thus

$$\mathbb{P}(u) \propto \exp\left(-\frac{1}{2}\langle u, Pu \rangle\right).$$

Likelihood For any $j \in Z'$

$$y(j) = S\left(u(j) + \eta(j)\right)$$

with the $\eta(j)$ drawn i.i.d from $\mathcal{N}(0, \gamma^2)$. We let

$$\Psi(v; \gamma) = \frac{1}{\sqrt{2\pi\gamma^2}} \int_{-\infty}^v \exp(-t^2/2\gamma^2) dt$$

and note that then

$$\mathbb{P}(y(j) = 1|u(j)) = \mathbb{P}(\mathcal{N}(0, \gamma^2) > -u(j)) = \Psi(u(j); \gamma) = \Psi(y(j)u(j); \gamma);$$

similarly

$$\mathbb{P}(y(j) = -1|u(j)) = \mathbb{P}(\mathcal{N}(0, \gamma^2) < -u(j)) = \Psi(-u(j); \gamma) = \Psi(y(j)u(j); \gamma).$$

Posterior Bayes' Theorem gives posterior μ_p with probability density function (pdf)

$$\mathbb{P}_p(u|y) \propto \exp\left(-\frac{1}{2}\langle u, Pu \rangle - \Phi_p(u; y)\right)$$

where

$$\Phi_p(u; y) := - \sum_{j \in Z'} \log(\Psi(y(j)u(j); \gamma)).$$

We let ν_p denote the push-forward under S of μ_p : $\nu_p = S\# \mu_p$.

MAP Estimator This is the minimizer of the negative of the log posterior. Thus we minimize the following objective function over U :

$$J_p(u) = \frac{1}{2}\langle u, Pu \rangle - \sum_{j \in Z'} \log\left(\Psi(y(j)u(j); \gamma)\right).$$

This is a convex function, a fact which is well-known in related contexts, but which we state and prove in the supplementary materials for the sake of completeness. In view of the close relationship between this problem and the level-set formulation described next, for which there are no minimizers, we expect that minimization may not be entirely straightforward in the $\gamma \ll 1$ limit. This is manifested in the presence of near-flat regions in the probit log likelihood function when $\gamma \ll 1$.

Our variant on the probit methodology differs from that in [20] in several ways: (i) our prior Gaussian is scaled to have per-node variance one, whilst in [20] the per node variance is a hyper-parameter to be determined; (ii) our prior is supported on $U = q_0^\perp$ whilst in [20] the prior precision is found by shifting L and taking a possibly fractional power of the resulting matrix, resulting in support on the whole of \mathbb{R}^N ; (iii) we allow for a scale parameter γ in the observational noise, whilst in [20] the parameter $\gamma = 1$.

3.2. Level-Set. This method is designed for problems considerably more general than classification on a graph [24]. For the current application, this model is exactly the same as probit except for the order in which the noise $\eta(j)$ and the thresholding function $S(u)$ is applied in the definition of the data.

Prior We again take as prior for u , the Gaussian μ_0 . Thus

$$\mathbb{P}(u) \propto \exp\left(-\frac{1}{2}\langle u, Pu \rangle\right).$$

Likelihood For any $j \in Z'$

$$y(j) = S(u(j)) + \eta(j)$$

with the $\eta(j)$ drawn i.i.d from $\mathcal{N}(0, \gamma^2)$. Then

$$\mathbb{P}(y(j)|u(j)) \propto \exp\left(-\frac{1}{2\gamma^2}|y(j) - S(u(j))|^2\right).$$

Posterior Bayes' Theorem gives posterior μ_{ls} with pdf

$$\mathbb{P}_{\text{ls}}(u|y) \propto \exp\left(-\frac{1}{2}\langle u, Pu \rangle - \Phi_{\text{ls}}(u; y)\right)$$

where

$$\Phi_{\text{ls}}(u; y) = \sum_{j \in Z'} \left(\frac{1}{2\gamma^2}|y(j) - S(u(j))|^2\right).$$

We let ν_{ls} denote the pushforward under S of μ_{ls} : $\nu_{\text{ls}} = S^\# \mu_{\text{ls}}$.

MAP Estimator Functional The negative of the log posterior is, in this case, given by

$$J_{\text{ls}}(u) = \frac{1}{2}\langle u, Pu \rangle + \Phi_{\text{ls}}(u; y).$$

However, unlike the probit model, the Bayesian level-set method has no MAP estimator – the infimum of J_{ls} is not attained and this may be seen by noting that, if the infimum was attained at any non-zero point u^* then ϵu^* would reduce the objective function for any $\epsilon \in (0, 1)$; however the point $u^* = 0$ does not attain the infimum. This proof is detailed in [24] for a closely related PDE based model, and the proof is easily adapted.

3.3. Atomic Noise Model. This is a variant on the level-set method, but deals with the fact that categorical data, whilst maybe noisy, will often be discrete. The resulting data model could also be used within the Ginzburg-Landau formulation in the next subsection, but we do not describe this explicitly. The prior we employ is the Gaussian μ_0 , and in contrast to the level-set method, but like the probit method, the observation is assumed to take values in the set $\{\pm 1\}$. In the atomic noise model the reporting of those values is accompanied by specified error rates [34], determined

by the parameters p, q . Taking $p = q = 1$ in the following corresponds to exact data with no error and is the same as level-set thresholding, or probit, in the limit $\gamma \rightarrow 0$.

Prior We again take as prior for u the Gaussian μ_0 . Thus

$$\mathbb{P}(u) \propto \exp\left(-\frac{1}{2}\langle u, Pu \rangle\right).$$

Likelihood We assume that the observation has the sign of u with a specified probability. To be precise we assume that we have,

$$\mathbb{P}(y = +1|u) = p, u \geq 0, \quad \mathbb{P}(y = +1|u) = 1 - q, u < 0$$

and

$$\mathbb{P}(y = -1|u) = 1 - p, u \geq 0, \quad \mathbb{P}(y = -1|u) = q, u < 0.$$

This defines a piecewise constant function of u , with discontinuity at $u = 0$, for each value $y \in \{\pm 1\}$: we call this function $\chi(u; y)$. In particular we have

$$\mathbb{P}(y(j)|u(j)) = \chi(u(j); y(j))$$

Posterior Bayes' Theorem gives posterior μ_{at} with pdf

$$\mathbb{P}_{\text{at}}(u|y) \propto \exp\left(-\frac{1}{2}\langle u, Pu \rangle - \Phi_{\text{at}}(u; y)\right).$$

where

$$\Phi_{\text{at}}(u; y) = - \sum_{j \in Z'} \log\left(\chi(u(j); y(j))\right).$$

We let ν_{at} denote the pushforward under S of $\mu_{\text{at}} : \nu_{\text{at}} = S\# \mu_{\text{at}}$.

MAP Estimator Functional This is the minimizer of the negative of the log posterior. Thus we minimize the objective function:

$$J_{\text{at}}(u) = \frac{1}{2}\langle u, Pu \rangle + \Phi_{\text{at}}(u; y).$$

We will not consider numerical methods for the atomic noise model because of space limitations. However it may be a natural choice for some measurement scenarios, hence its inclusion in this paper. Its behaviour is very similar to the Bayesian level set method because its likelihood is also piecewise constant with respect to latent variable u ; and as with the Bayesian level set there is no minimizer for the MAP estimation problem.

3.4. Ginzburg-Landau. For this model, we take as prior the Ginzburg-Landau measure ν_0 defined by (6), and employ a Gaussian likelihood for the observed labels. This construction gives the Bayesian posterior whose MAP estimator is the objective function introduced and studied in [5].

Prior We define prior on v to be the Ginzburg-Landau measure ν_0 given by (6) with density

$$\mathbb{P}(v) \propto e^{-\text{GL}(v)}.$$

Likelihood For any $j \in Z'$

$$y(j) = v(j) + \eta(j)$$

with the $\eta(j)$ drawn i.i.d from $\mathcal{N}(0, \gamma^2)$. Then

$$\mathbb{P}\left(y(j)|v(j)\right) \propto \exp\left(-\frac{1}{2\gamma^2}|y(j) - v(j)|^2\right).$$

Posterior Recalling that $P = c^{-1}L$ we see that Bayes' Theorem gives posterior ν_{gl} with pdf

$$\begin{aligned} \mathbb{P}_{\text{gl}}(v|y) &\propto \exp\left(-\frac{1}{2}\langle v, Pv \rangle - \Phi_{\text{gl}}(v; y)\right), \\ \Phi_{\text{gl}}(v; y) &:= \sum_{j \in Z} W_{\epsilon}(v(j)) + \sum_{j \in Z'} \left(\frac{1}{2\gamma^2}|y(j) - v(j)|^2\right). \end{aligned}$$

MAP Estimator This is the minimizer of the negative of the log posterior. Thus we minimize the following objective function over U :

$$J_{\text{gl}}(v) = \frac{1}{2}\langle v, Pv \rangle + \Phi_{\text{gl}}(v; y).$$

This objective function was introduced in [5] as a relaxation of the min-cut problem, penalized by data; the relationship to min-cut was studied rigorously in [40]. The minimization problem for J_{gl} is non-convex and has multiple minimizers, reflecting the combinatorial character of the min-cut problem of which it is a relaxation.

3.5. Small Label Noise Limit. In the small label noise limit $\gamma = 0$ the probit and level-set posteriors coincide with the atomic noise model in the limit where $p = q = 1$. Furthermore all models then take the form of the Gaussian prior μ_0 conditioned to be positive on labelled nodes where $y(j) = 1$ and to be negative on labelled nodes where $y(j) = -1$. This can be linked with the original work of Zhu et al [49, 50] which based classification on the measure μ_0 conditioned to take the value exactly 1 on labelled nodes where $y(j) = 1$ and conditioned to take the value exactly -1 on labelled nodes where $y(j) = -1$. Thus we see explicit connections between a variety of different Bayesian formulations of graph-based semi-supervised learning.

3.6. Uncertainty Quantification for Graph Based Learning. In Figure 1 we plot the component of the negative log likelihood at a labelled node j , as a function of the latent variable $u = u(j)$ with data $y = y(j)$ fixed, for the probit, Bayesian level-set, and atomic noise models. The log likelihood for the Ginzburg-Landau formulation is not directly comparable as it is a function of the relaxed label variable $v(j)$, with respect to which it is quadratic with minimum at the data point $y(j)$.

The probit, Bayesian level-set, and atomic noise models lead to posterior distributions μ (with different subscripts) in latent variable space, and pushforwards ν (also with different subscripts) in label space. The Ginzburg-Landau formulation leads to a measure ν_{gl} in label space. Uncertainty quantification in the widest sense is concerned with completely characterizing these posterior distributions. In practice this may be achieved by sampling using MCMC methods. In this paper we will study four measures of uncertainty:

- we will study the empirical pdfs of the latent and label variables at certain nodes;
- we will study the posterior mean of the label variables at certain nodes;
- we will study the posterior variance of the label variables averaged over all nodes;
- we will use the posterior mean or variance to order nodes into those whose classifications are most uncertain and those which are most certain.

For the probit, level-set and atomic models, we interpret the thresholded variable $l = S(u)$ as the binary label assignments corresponding to a real-valued configuration u . The node-wise posterior mean of l can be used as a useful confidence score of the

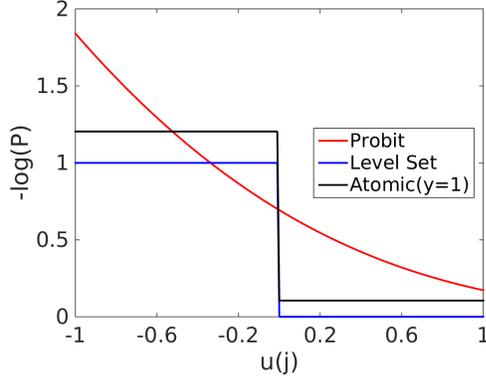


FIG. 1. Plot of a component of the negative log likelihood for a fixed node j . We set $\gamma = 1/\sqrt{2}$ for probit and Bayesian level-set, and $p = 0.8$, $q = 0.7$ for atomic noise model. Since $\Phi(u(j); 1) = \Phi(-u(j); -1)$ for probit and Bayesian level-set, we omit the plot for $y(j) = -1$.

class assignment of each node. The node-wise posterior mean s_j^l is defined as

$$(8) \quad s_j^l := \mathbb{E}_\nu(l(j)),$$

with respect to any of the posterior measures (pushed forward from latent variable space for probit, level-set and atomic models) ν . Note that $s_j^l \in [-1, 1]$ and if $q = \nu(l(j) = 1)$ then $q = \frac{1}{2}(1 + s_j^l)$. For binary labels $l(j) \in \{\pm 1\}$ the mean also contains the variance information, and hence the formula (8) captures posterior variance. Specifically we have that

$$\text{Var}_\nu(l(j)) = 4q(1 - q) = 1 - (s_j^l)^2.$$

Later we will find it useful to consider the variance averaged over all nodes and hence define³

$$(9) \quad \text{Var}(l) = \frac{1}{N} \sum_{j=1}^N \text{Var}_\nu(l(j)).$$

Note that the maximum value obtained by $\text{Var}(l)$ is 1. This maximum value is attained under all the prior distributions we use in this paper. The deviation from this maximum, under the posterior, is a measure of the information content of the labelled data. Note, however, that the prior does contain information about classifications, in the form of correlations between vertices; this is not captured in (9).

4. Algorithms. From Section 3, we see that for all of the models considered, the posterior $\mathbb{P}(w|y)$ has the form

$$\mathbb{P}(w|y) \propto \exp(-J(w)), \quad J(w) = \frac{1}{2} \langle w, Pw \rangle + \Phi(w)$$

for some function Φ , different for each of the four models (acknowledging that in the Ginzburg-Landau case the independent variable is $w = v$, real-valued relaxation of

³Strictly speaking $\text{Var}(l) = N^{-1} \text{Tr}(\text{Cov}(l))$.

label space, where as for the other models $w = u$ an underlying latent variable which may be thresholded by $S(\cdot)$ into label space.) Furthermore, the MAP estimator is the minimizer of J . Note that Φ is differentiable for the Ginzburg-Landau and probit models, but not for the level-set and atomic noise models. We introduce algorithms for both sampling (MCMC) and MAP estimation (optimization) that apply in this general framework. The sampler we employ does not use information about the gradient of Φ ; the MAP estimation algorithm does, but is only employed on the Ginzburg-Landau and probit models. The samplers do use properties of the precision matrix P , which is proportional to the graph Laplacian L ; in particular its spectral properties are relevant. Figure 2 demonstrates the spectral properties of L for the four examples that we will apply our algorithms to in section 5.

4.1. MCMC. We sample the posterior probability distribution using MCMC. To date probit models have typically been sampled by means of a Gibbs methodology.

However for three reasons we consider sampling algorithms which apply directly on all the nodes Z . These are: (i) we wish to highlight methods which apply to the Ginzburg-Landau, level-set and atomic noise models which precludes the explicit conditionally Gaussian, or truncated Gaussian, form of the methods described for probit; (ii) all of our posterior distributions have a density with respect to the Gaussian μ_0 – that is their densities are proportional to that of μ_0 – and as a result we may use MCMC methods which, in the case where the graph Laplacian has a limit [39], have the potential for delivering samples from the posterior in a number of steps which is independent of the dimension N of the state space, as overviewed in [15]; (iii) these MCMC methods are well-adapted to the use of approximation methods which exploit structure in the spectral properties of the graph Laplacian. Other classes of MCMC methods, such as the Gibbs samplers in [1, 22]; could be considered; and other priors, relaxing the Gaussian structure, could be considered; but in taking these directions then the development of MCMC methods with N –independent mixing rates which can also exploit approximations of the spectral properties of the graph Laplacian is an open research direction.

In order to induce scalability with respect to size of Z we use the pCN method described in [15] and introduced in the context of diffusions by Beskos et. al. in [7] and by Neal in the context of machine learning [31]. The standard random walk Metropolis (RWM) algorithm suffers from the fact that the optimal proposal variance or stepsize scales inverse proportionally to the dimension of the state space [35], which is the graph size N in this case. The pCN method is designed so that the proposal variance required to obtain a given acceptance probability scales independently of the dimension of the state space (here the number of graph nodes N), hence in practice giving faster convergence of the MCMC when compared with RWM [6]. We restate the pCN method as Algorithm 1, and then follow with various variants on it in Algorithms 2 and 3. In all three algorithms $\beta \in [0, 1]$ is the key parameter which determines the efficiency of the MCMC method: small β leads to high acceptance probability but small moves; large β leads to low acceptance probability and large moves. Somewhere between these extremes is an optimal choice of β which minimizes the asymptotic variance of the algorithm when applied to compute a given expectation.

The value $\xi^{(k)}$ is a sample from the prior μ_0 . If the eigenvalues and eigenvectors of L are all known then the Karhunen-Loeve expansion (10) gives

$$(10) \quad \xi^{(k)} = c^{\frac{1}{2}} \sum_{j=1}^{N-1} \lambda_j^{-\frac{1}{2}} q_j z_j,$$

Algorithm 1 pCN Algorithm

- 1: Input: L . $\Phi(u)$. $u^{(0)} \in U$.
 - 2: Output: M Approximate samples from the posterior distribution
 - 3: Define: $\alpha(u, w) = \min\{1, \exp(\Phi(u) - \Phi(w))\}$.
 - 4: **while** $k < M$ **do**
 - 5: $w^{(k)} = \sqrt{1 - \beta^2}u^{(k)} + \beta\xi^{(k)}$, where $\xi^{(k)} \sim \mathcal{N}(0, C)$ via Eq.(10).
 - 6: Calculate acceptance probability $\alpha(u^{(k)}, w^{(k)})$.
 - 7: Accept $w^{(k)}$ as $u^{(k+1)}$ with probability $\alpha(u^{(k)}, w^{(k)})$, otherwise $u^{(k+1)} = u^{(k)}$.
 - 8: **end while**
-

where c is given by (5), the $z_j, j = 1 \dots N - 1$ are i.i.d centred unit Gaussians and the equality is in law.

4.2. Spectral Projection. For graphs with a large number of nodes N , it is prohibitively costly to directly sample from the distribution μ_0 , since doing so involves knowledge of a complete eigen-decomposition of L , in order to employ (10). A method that is frequently used in classification tasks is to restrict the support of u to the eigenspace spanned by the first ℓ eigenvectors with the smallest non-zero eigenvalues of L (hence largest precision) and this idea may be used to approximate the pCN method; this leads to a low rank approximation. In particular we approximate samples from μ_0 by

$$(11) \quad \xi_\ell^{(k)} = c_\ell^{\frac{1}{2}} \sum_{j=1}^{\ell-1} \lambda_j^{-\frac{1}{2}} q_j z_j,$$

where c_ℓ is given by (5) truncated after $j = \ell - 1$, the z_j are i.i.d centred unit Gaussians and the equality is in law. This is a sample from $\mathcal{N}(0, C_\ell)$ where $C_\ell = c_\ell Q \Sigma_\ell Q^*$ and the diagonal entries of Σ_ℓ are set to zero for the entries after ℓ . In practice, to implement this algorithm, it is only necessary to compute the first ℓ eigenvectors of the graph Laplacian L . This gives Algorithm 2.

Algorithm 2 pCN Algorithm With Spectral Projection

- 1: Input: L . $\Phi(u)$. $u^{(0)} \in U$.
 - 2: Output: M Approximate samples from the posterior distribution
 - 3: Define: $\alpha(u, w) = \min\{1, \exp(\Phi(u) - \Phi(w))\}$.
 - 4: **while** $k < M$ **do**
 - 5: $w^{(k)} = \sqrt{1 - \beta^2}u^{(k)} + \beta\xi_\ell^{(k)}$, where $\xi_\ell^{(k)} \sim \mathcal{N}(0, C_\ell)$ via Eq.(11).
 - 6: Calculate acceptance probability $\alpha(u^{(k)}, w^{(k)})$.
 - 7: Accept $w^{(k)}$ as $u^{(k+1)}$ with probability $\alpha(u^{(k)}, w^{(k)})$, otherwise $u^{(k+1)} = u^{(k)}$.
 - 8: **end while**
 - 9: **return** u_k
-

The accuracy of Algorithm 2 as an approximation of Algorithm 1 depends to a large extent on the size of the eigenvalues of L in the following sense:

$$(12) \quad \mathbb{E}|\xi^{(k)}|^2 - \mathbb{E}|\xi_\ell^{(k)}|^2 = c \sum_{j=1}^{N-1} \lambda_j^{-1} - c_\ell \sum_{j=1}^{\ell-1} \lambda_j^{-1}$$

where the expectation is with respect to the centred unit Gaussians $\{z_j\}$. Note that the examples shown in Figure 2 gives an indication of the quality of this approximation;

the size and number of the smallest eigenvalues of L play a very important role in determining whether the difference in (12) is small.

4.3. Spectral Approximation. Spectral projection often leads to good classification results, but may lead to reduced posterior variance and a posterior distribution that is overly smooth on the graph domain. We propose an improvement on the method that preserves the variability of the posterior distribution but still only involves calculating the first ℓ eigenvectors of L . This is based on the empirical observation that in many applications the spectrum of L saturates and satisfies, for $j \geq \ell$, $\lambda_j \approx \bar{\lambda}$ for some $\bar{\lambda}$. Such behaviour may be observed in b), c) and d) of Figure 2; in particular note that in the hyperspectral case $\ell \ll N$. We assume such behaviour in deriving the low rank approximation used in this subsection. (See supplementary materials for a detailed discussion of the graph Laplacian spectrum.) We define $\Sigma_{\ell,o}$ by overwriting the diagonal entries from ℓ to $N-1$ with $\bar{\lambda}^{-1}$. We then set $C_{\ell,o} = c_{\ell,o} Q \Sigma_{\ell,o} Q^*$, and generate approximate samples from μ_0 by setting

$$(13) \quad \xi_{\ell,o}^{(k)} = c_{\ell,o}^{\frac{1}{2}} \sum_{j=1}^{\ell-1} \lambda_j^{-\frac{1}{2}} q_j z_j + c_{\ell,o}^{\frac{1}{2}} \bar{\lambda}^{-\frac{1}{2}} \sum_{j=\ell}^{N-1} q_j z_j,$$

where $c_{\ell,o}$ is given by (5) with λ_j replaced by $\bar{\lambda}$ for $j \geq \ell$, the $\{z_j\}$ are centred unit Gaussians, and the equality is in law. Importantly samples according to (13) can be computed very efficiently. In particular there is no need to compute q_j for $j \geq \ell$, and the quantity $\sum_{j=\ell}^{N-1} q_j z_j$ can be computed by first taking a sample $\bar{z} \sim \mathcal{N}(0, I_N)$, and then projecting \bar{z} onto $U_\ell := \text{span}(q_\ell, \dots, q_{N-1})$. Moreover, projection onto U_ℓ can be computed only using $\{q_1, \dots, q_{\ell-1}\}$, since the vectors span the orthogonal complement of U_ℓ . Concretely, we have

$$\sum_{j=\ell}^{N-1} q_j z_j = \bar{z} - \sum_{j=1}^{\ell-1} q_j \langle q_j, \bar{z} \rangle,$$

where $\bar{z} \sim \mathcal{N}(0, I_N)$ and equality is in law. Hence the samples $\xi_{\ell,o}^{(k)}$ can be computed by

$$(14) \quad \xi_{\ell,o}^{(k)} = c_{\ell,o}^{\frac{1}{2}} \sum_{j=1}^{\ell-1} \lambda_j^{-\frac{1}{2}} q_j z_j + c_{\ell,o}^{\frac{1}{2}} \bar{\lambda}^{-\frac{1}{2}} \left(\bar{z} - \sum_{j=1}^{\ell-1} q_j \langle q_j, \bar{z} \rangle \right).$$

The error induced by this approximation can be characterized through the formula

$$(15) \quad \mathbb{E}|\xi^{(k)}|^2 - \mathbb{E}|\xi_{\ell,o}^{(k)}|^2 = (c - c_{\ell,o}) \sum_{j=1}^{N-1} \lambda_j^{-1} + c_{\ell,o} \sum_{j=\ell}^{N-1} (\lambda_j^{-1} - \bar{\lambda}^{-1}).$$

Under the stated empirical properties of the graph Laplacian, we expect this to be a better approximation of the prior variance than the approximation leading to (12). The vector $\xi_{\ell,o}^{(k)}$ is a sample from $\mathcal{N}(0, C_{\ell,o})$ and results in Algorithm 3.

4.4. MAP Estimation: Optimization. Recall that the objective function for the MAP estimation has the form $\frac{1}{2} \langle u, Pu \rangle + \Phi(u)$, where u is supported on the space U . For Ginzburg-Landau and probit, the function Φ is smooth, and we can use a standard projected gradient method for the optimization. Since L is typically

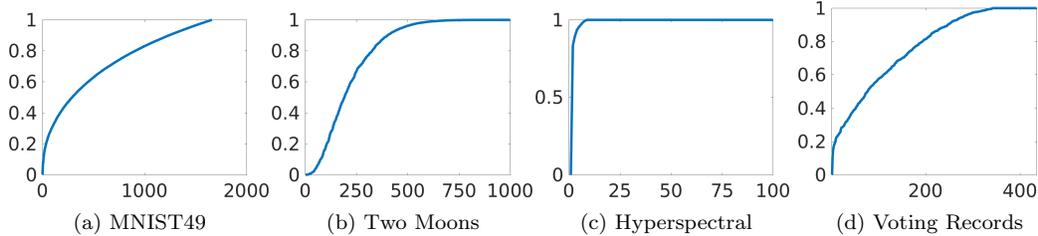


FIG. 2. Spectra of graph Laplacian of various datasets. See Sec.5 for the description of the datasets and graph construction parameters.

Algorithm 3 pCN Algorithm With Spectral Approximation

- 1: Input: L , $\Phi(u)$, $u^{(0)} \in U$.
 - 2: Output: M Approximate samples from the posterior distribution
 - 3: Define: $\alpha(u, w) = \min\{1, \exp(\Phi(u) - \Phi(w))\}$.
 - 4: **while** $k < M$ **do**
 - 5: $w^{(k)} = \sqrt{1 - \beta^2}u^{(k)} + \beta\xi_{\ell,o}^{(k)}$, where $\xi_{\ell,o}^{(k)} \sim \mathcal{N}(0, C_{\ell,o})$ via Eq.(14).
 - 6: Calculate acceptance probability $\alpha(u^{(k)}, w^{(k)})$.
 - 7: Accept $w^{(k)}$ as $u^{(k+1)}$ with probability $\alpha(u^{(k)}, w^{(k)})$, otherwise $u^{(k+1)} = u^{(k)}$.
 - 8: **end while**
 - 9: **return** u_k
-

ill-conditioned, it is preferable to use a semi-implicit discretization as suggested in [5], as convergence to a stationary point can be shown under a graph independent learning rate. Furthermore, the discretization can be performed in terms of the eigenbasis $\{q_1, \dots, q_{N-1}\}$, which allows us to easily apply spectral projection when only a truncated set of eigenvectors is available. We state the algorithm in terms of the (possibly truncated) eigenbasis below. Here P_ℓ is an approximation to P found by setting $P_\ell = Q_\ell D_\ell Q_\ell^*$ where Q_ℓ is the matrix with columns $\{q_1, \dots, q_{\ell-1}\}$ and $D_\ell = \text{diag}(d)$ for $d(j) = c_\ell \lambda_j$, $j = 1, \dots, \ell - 1$. Thus $P_{N-1} = P$.

Algorithm 4 Linearly-Implicit Gradient Flow with Spectral Projection

- 1: **Input:** $Q_m = (q_1, \dots, q_m)$, $\Lambda_m = (\lambda_1, \dots, \lambda_m)$, $\Phi(u)$, $u^{(0)} \in U$.
 - 2: **while** $k < M$ **do**
 - 3: $u^{(*)} = u^{(k)} - \beta \nabla \Phi(u^{(k)})$
 - 4: $u^{(k+1)} = (I + \beta P_m)^{-1} u^{(*)}$
 - 5: **end while**
-

5. Numerical Experiments. In this section we conduct a series of numerical experiments on four different data sets that are representative of the field of graph semi-supervised learning. There are four main purposes for the experiments. First we perform uncertainty quantification, as explained in subsection 3.6. Secondly, we study the spectral approximation and projection variants on pCN sampling as these scale well to massive graphs. Finally we make some observations about the cost and practical implementation details of these methods, for the different Bayesian models we adopt; these will help guide the reader in making choices about which algorithm to use. We present the results for MAP estimations in the supplementary materials,

alongside the proof of convexity of the Probit MAP estimator.

The quality of the graph constructed from the feature vectors is central to the performance of any graph learning algorithms. In the experiments below, we follow the graph construction procedures used in the previous papers [5, 23, 29]; those papers, together, applied graph partitioning to all of the datasets that we use here and so provide important guidance. Moreover, we have verified that for all the reported experiments below, the graph parameters are in a range such that spectral clustering gives a reasonable performance. The methods we employ lead to refinements over spectral clustering (improved classification) and, of course, to uncertainty quantification (which spectral clustering does not address).

5.1. Data Sets. We introduce the data sets and describe the graph construction for each data set. In all cases we numerically construct the weight matrix A , and then the graph Laplacian L .⁴

5.1.1. Two Moons. The two moons artificial data set is constructed to give noisy data which lies near a nonlinear low dimensional manifold embedded in a high dimensional space [13]. The data set is constructed by sampling N data points uniformly from two semi-circles centered at $(0, 0)$ and $(1, 0.5)$ with radius 1, embedding the data in \mathbb{R}^d , and adding Gaussian noise with standard deviation σ . We set $N = 2,000$ and $d = 100$ in this paper; recall that then the graph size is N and each feature vector has length d . We will conduct a variety of experiments with different labelled data size J , and in particular study variation with J . The default value, when not varied, is J at 3% of N , with the labelled points chosen at random.

We take each data point as a node on the graph, and construct a fully connected graph using the self-tuning weights of Zelnik-Manor and Perona [46], with $K = 10$. Specifically we let x_i, x_j be the coordinates of the data points i and j . Then weight w_{ij} from i to j is defined by

$$(16) \quad w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\tau_i\tau_j}\right),$$

where τ_j is the distance of the K -th closest point to the node j .

5.1.2. House Voting Records from 1984. This dataset contains the voting records of 435 U.S. House of Representatives; for details see [5] and the references therein. The votes were recorded in 1984 from the 98th United States Congress, 2nd session. The votes for each individual is vectorized by mapping a yes vote to 1, a no vote to -1 , and an abstention/no-show to 0. The data set contains 16 votes that are believed to be well-correlated with partisanship, and we use only these votes as feature vectors for constructing the graph. Thus the graph size is $N = 435$, and feature vectors have length $d = 16$. The goal is to predict the party affiliation of each individual. We pick 3 Democrats and 2 Republicans at random to use as the observed class labels; thus $J = 5$ corresponding to less than 1.2% of fidelity points. We construct a fully connected graph with weights given by (16) with $\tau_j = \tau = 1.25$ for all nodes j .

5.1.3. MNIST. The MNIST database consists of 70,000 images of size 28×28 pixels containing the handwritten digits 0 through 9; see [26] for details. Since in this paper we focus on binary classification, we only consider pairs of digits. To speed up

⁴The weight matrix A is symmetric in theory; in practice we find that symmetrizing via the map $A \mapsto \frac{1}{2}A + \frac{1}{2}A^*$ is helpful.

calculations, we subsample randomly 2,000 images from each digit to form a graph with $N = 4,000$ nodes; we use this for all our experiments except in subsection 5.4 where we use the full data set of size $N = \mathcal{O}(10^4)$ for digit pair (4, 9) to benchmark computational cost. The nodes of the graph are the images and as feature vectors we project the images onto the leading 50 principal components given by PCA; thus the feature vectors at each node have length $d = 50$. We conduct a variety of experiments with different labelled data dimension J , and in particular study variation with J . The default value, when not varied, is J at 4% of N , with the labelled points chosen at random. We construct a K -nearest neighbor graph with $K = 20$ for each pair of digits considered. Namely, the weights A_{ij} are non-zero if and only if one of i or j is in the K nearest neighbors of the other. The non-zero weights are set using (16) with $K = 20$. This is the only example we consider in this paper that does not have a fully connected graph.

We choose the four pairs (5, 7), (0, 6), (3, 8) and (4, 9). These four pairs exhibit increasing levels of difficulty for classification. This fact is demonstrated in Figures 3a - 3d, where we visualize the datasets by projecting the dataset onto the second and third eigenvector of the graph Laplacian. Namely, each node i is mapped to the point $(Q(2, i), Q(3, i)) \in \mathbb{R}^2$, where $L = Q\Lambda Q^*$.

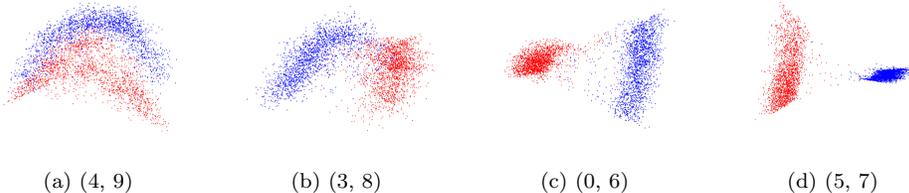


FIG. 3. Visualization of data by projection onto 2^{nd} and 3^{rd} eigenfunctions of the graph Laplacian for the MNIST data set, where the vertical dimension is the 3^{rd} eigenvector and the horizontal dimension the 2^{nd} . Each subfigure represents a different pair of digits. We construct a 20 nearest neighbour graph under the Zelnik-Manor and Perona scaling [46] as in (16) with $K = 20$.

5.1.4. HyperSpectral Image. The hyperspectral data set analysed for this project was provided by the Applied Physics Laboratory at Johns Hopkins University; see [12] for details. It consists of a series of video sequences recording the release of chemical plumes taken at the Dugway Proving Ground. Each layer in the spectral dimension depicts a particular frequency starting at 7,830 nm and ending with 11,700 nm, with a channel spacing of 30 nm, giving 129 channels; thus the feature vector has length $d = 129$. The spatial dimension of each frame is 128×320 pixels. We select 7 frames from the video sequence as the input data, and consider each spatial pixel as a node on the graph. Thus the graph size is $N = 128 \times 320 \times 7 = 286,720$. The classification problem is to classify pixels that represent the chemical plumes against pixels that are the background.

We construct a fully connected graph with weights given by the cosine distance:

$$w_{ij} = \frac{\langle x_i, x_j \rangle}{\|x_i\| \|x_j\|}.$$

This distance is small for vectors that point in the same direction, and is insensitive to their magnitude. We consider the symmetric Laplacian defined in (1). Because

it is computationally prohibitive to compute eigenvectors of a Laplacian of this size, we apply the Nyström extension [44, 17] to obtain an approximation to the true eigenvectors and eigenvalues; see [5] for details pertinent to the set-up here. We emphasize that each pixel in the 7 frames is a node on the graph and that, in particular, pixels across the 7 time-frames are also connected. Since we have no ground truth labels for this dataset, we generate known labels by setting the segmentation results from spectral clustering as ground truth. The default value of J is 8,000, and labels are chosen at random. This corresponds to labelling around 2.8% of the points. We only plot results for the last 6 frames of the video sequence since the first frame does not contain the chemical plume.

5.2. Uncertainty Quantification. In this subsection we demonstrate both the feasibility, and value, of uncertainty quantification in graph classification methods. We employ the probit and the Bayesian level-set model for most of the experiments in this subsection; we also employ the Ginzburg-Landau model but since this can be slow to converge, due to the presence of local minima, it is only demonstrated on the voting records dataset. The atomic noise model has a similar piecewise constant log likelihood as for the Bayesian level-set method, and so we omit experiments on the atomic noise model. The pCN method is used for sampling on various datasets to demonstrate properties and interpretations of the posterior.

5.2.1. Visualization of Marginal Posterior Density. In this subsection, we contrast the posterior distribution $\mathbb{P}(v|y)$ of the Ginzburg-Landau model with that of the probit and Bayesian level-set (BLS) models. The graph is constructed from the voting records data with the fidelity points chosen as described in subsection 5.1. In Figure 4 we plot the histograms of the empirical marginal posterior distribution on $\mathbb{P}(v(i)|y)$ and $\mathbb{P}(u(i)|y)$ for a selection of nodes on the graph. For the top row of Figure 4, we select 6 nodes with “low confidence” predictions, and plot the empirical marginal distribution of u for probit and BLS, and that of v for the Ginzburg-Landau model. Note that the same set of nodes is chosen for different models. The plots in this row demonstrate the multi-modal nature of the Ginzburg-Landau distribution in contrast to the uni-modal nature of the probit posterior; this uni-modality is a consequence of Proposition 1. For the bottom row, we plot the same empirical distributions for 6 nodes with “high confidence” predictions. In contrast with the top row, the Ginzburg-Landau marginal for high confidence nodes is essentially uni-modal since most samples of v evaluated on these nodes have a fixed sign.

We also observe that the pCN algorithm converges far more quickly for probit than for Ginzburg-Landau, because of the presence of multiple modes in the latter; this is manifest in the fact that the posteriors for Ginzburg-Landau are less well converged than for probit, for a similar amount of algorithmic time; this issue is quantified in subsection 5.4. This undesirable feature of Ginzburg-Landau sampling can be ameliorated by choosing a larger value of ϵ ; however this then leads to a probability measure which is further from the label space which it relaxes. The Bayesian level-set method behaves similarly to probit as is to be expected from the fact that, in the limit $\gamma \rightarrow 0$, they are formally identical as discussed in subsection 3.5.

5.2.2. Posterior Mean as Confidence Scores. We construct the graph from the MNIST (4,9) dataset following subsection 5.1. The noise variance γ is set to 0.1, and 4% of fidelity points are chosen randomly from each class. The probit posterior is used to compute (8). In Figure 5 we demonstrate that nodes with scores s_j^l closer to the binary ground truth labels ± 1 look visually more uniform than nodes with s_j^l

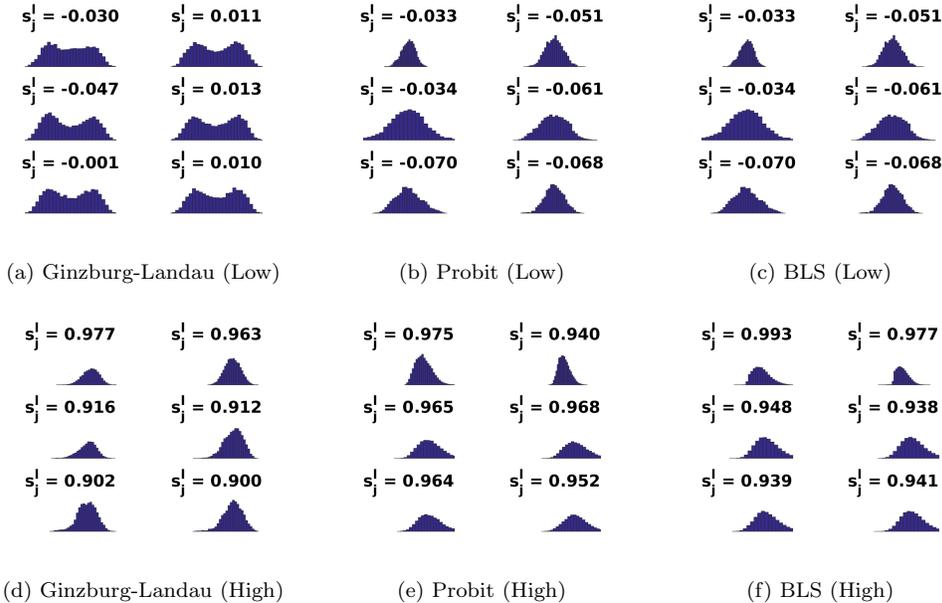


FIG. 4. Visualization of marginal posterior density for low and high confidence predictions across different models. Each image plots the empirical marginal posterior density of a certain node i , obtained from the histogram of 1×10^5 approximate samples using pCN. Columns in the figure (e.g. a) and d)) are grouped by model. From left to right, the models are Ginzburg-Landau, probit, and Bayesian level-set respectively. From the top down, the rows in the figure (e.g. a)-c)) denote the low confidence and high confidence predictions respectively. For the top row, we select 6 nodes with the lowest absolute value of the posterior mean s_j^l , defined in Eq.(8), averaged across three models. Note that the same set of nodes is chosen for different models. These nodes represent outliers in the dataset that are hard to classify, and hence more likely to induce a multi-modal marginal in the Ginzburg-Landau model. For the bottom row, we select nodes with the highest average posterior mean s_j^l . These nodes represent nodes that are classified with greatest certainty to the class label +1. We present the posterior mean s_j^l on top of the histograms for reference. The experiment parameters are: $\epsilon = 10.0$, $\gamma = 0.6$, $\beta = 0.1$ for the Ginzburg-Landau model, and $\gamma = 0.5$, $\beta = 0.2$ for the probit and BLS model.

far from those labels. This shows that the posterior mean contains useful information which differentiates between outliers and inliers that align with human perception.

5.2.3. Posterior Variance as Uncertainty Measure. In this set of experiments, we show that the posterior distribution of the label variable $l = S(u)$ captures the uncertainty of the classification problem. We use the posterior variance of l , averaged over all nodes, as a measure of the model variance; specifically formula (9). We study the behaviour of this quantity as we vary the level of uncertainty within certain inputs to the problem. We demonstrate empirically that the posterior variance is approximately monotonic with respect to variations in the levels of uncertainty in the input data, as it should be; and thus that the posterior variance contains useful information about the classification. We select quantities that reflect the separability of the classes in the feature space.

Figure 6 plots the posterior variance $\text{Var}(l)$ against the standard deviation σ of the noise appearing in the feature vectors for the two moons dataset; thus points generated on the two semi-circles overlap more as σ increases. We employ a sequence of posterior

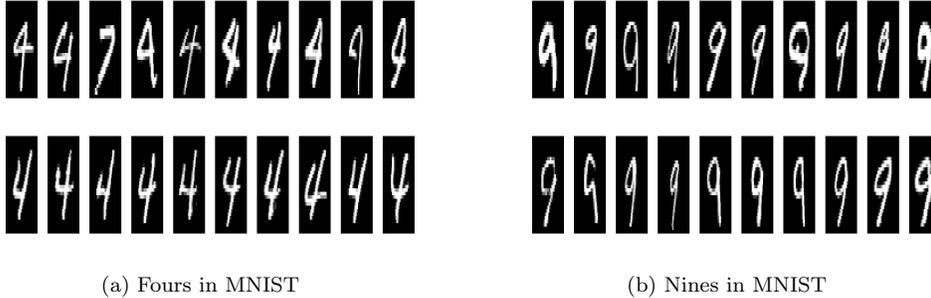


FIG. 5. “Hard to classify” vs “easy to classify” nodes in the MNIST (4,9) dataset under the probit model. Here the digit “4” is labeled +1 and “9” is labeled -1. The top (bottom) row of the left column corresponds to images that have the lowest (highest) values of s_j^l defined in (8) among images that have ground truth labels “4”. The right column is organized in the same way for images with ground truth labels 9 except the top row now corresponds to the highest values of s_j^l . Higher s_j^l indicates higher confidence that image j is a 4 and not a “9”, hence the top row could be interpreted as images that are “hard to classify” by the current model, and vice versa for the bottom row. The graph is constructed as in Section 5, and $\gamma = 0.1$, $\beta = 0.3$.

computations, using probit and Bayesian level-set, for $\sigma = 0.02 : 0.01 : 0.12$. Recall that $N = 2,000$ and we choose 3% of the nodes to have the ground truth labels as observed data. Within both models, γ is fixed at 0.1. A total of 1×10^4 samples are taken, and the proposal variance β is set to 0.3. We see that the mean posterior variance increases with σ , as is intuitively reasonable. Furthermore, because γ is small, probit and Bayesian level-set are very similar models and this is reflected in the similar quantitative values for uncertainty.

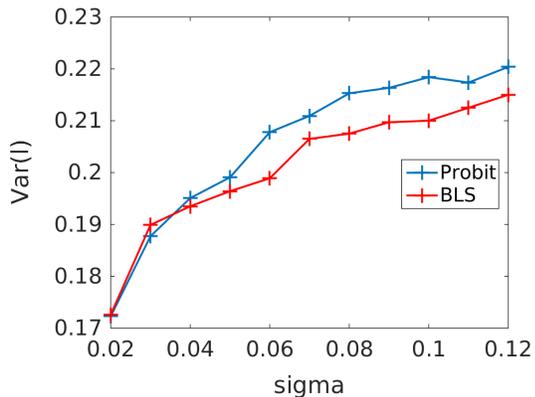


FIG. 6. Mean Posterior Variance defined in (9) versus feature noise σ for the probit model and the BLS model applied to the Two Moons Dataset with $N = 2,000$. For each trial, a realization of the two moons dataset under the given parameter σ is generated, and 3% of nodes are randomly chosen as fidelity. We run 20 trials for each value of σ , and average the mean posterior variance across the 20 trials in the figure. We set $\gamma = 0.1$ and $\beta = 0.3$ for both models.

A similar experiment studies the posterior label variance $\text{Var}(l)$ as a function of the pair of digits classified within the MNIST data set. We choose 4% of the nodes as labelled data, and set $\gamma = 0.1$. The number of samples employed is 1×10^4 and the

proposal variance β is set to be 0.3. Table 1 shows the posterior label variance. Recall that Figures 3a - 3d suggest that the pairs (4, 9), (3, 8), (0, 6), (5, 7) are increasingly easy to separate, and this is reflected in the decrease of the posterior label variance shown in Table 1.

Digits	(4, 9)	(3, 8)	(0, 6)	(5, 7)
probit	0.1485	0.1005	0.0429	0.0084
BLS	0.1280	0.1018	0.0489	0.0121

TABLE 1

Mean Posterior Variance of different digit pairs for the probit model and the BLS model applied to the MNIST Dataset. The pairs are organized from left to right according to the separability of the two classes as shown in Fig.3a - 3d. For each trial, we randomly select 4% of nodes as fidelity. We run 10 trials for each pairs of digits and average the mean posterior variance across trials. We set $\gamma = 0.1$ and $\beta = 0.3$ for both models.

The previous two experiments in this subsection have studied posterior label variance $\text{Var}(l)$ as a function of variation in the prior data; specifically as a function of the noise defining the feature vectors in two moons, and on the pair of digits used to classify in MNIST. We now turn and study how posterior variance changes as a function of varying the likelihood information, again for both two moons and MNIST data sets. In the two moons data set we freeze the feature vector noise at $\sigma = 0.06$. In Figures 7a and 7b, for the two moons and MNIST (4, 9) data sets respectively, we plot the posterior label variance against the percentage of nodes observed. We observe that the observational variance decreases as the amount of labelled data increases. Figures 8a and 8b plot the posterior label variance as the observational noise γ is varied in the probit model, for both the two moons and MNIST (4, 9) data sets; we fix 3% and 4% of randomly chosen nodes as observed labels in parts (a) and (b) of the figure respectively. Note that, as the observational variance increases, so too does the posterior label variance. Furthermore the level set and probit formulations produce similar answers for γ small, reflecting the discussion in subsection 3.5.

In summary of this subsection, the label posterior variance $\text{Var}(l)$ behaves intuitively as expected as a function of varying the prior and likelihood information that specify the statistical probit model and the Bayesian level-set model; furthermore these two Bayesian models produce quantitatively similar results because they are formally identical when noise $\gamma \rightarrow 0$. The uncertainty quantification thus provides useful, and consistent, information that can be used to inform decisions made on the basis of classifications.

5.3. Spectral Approximation and Projection Methods. Here we discuss Algorithms 2 and 3, designed to approximate the full (but expensive on large graphs) Algorithm 1. In the first subsection we consider the voting records problem. This is small enough to compare the posterior distribution obtained from spectral projection and approximation with full sampling, and thereby verify their properties. Armed with this information we then study the hyperspectral problem in the next subsection; this problem is too large to be amenable to full sampling.

5.3.1. Applications to Voting Records. We study how the spectral projection and approximation methods, Algorithm 2 and Algorithm 3, compare in performance with the full posterior distribution sampled via Algorithm 1. We do this by comparing the posterior mean of the thresholded variable s_j^l in each case, and the results are shown in Figure 9. This clearly demonstrates that spectral projection does

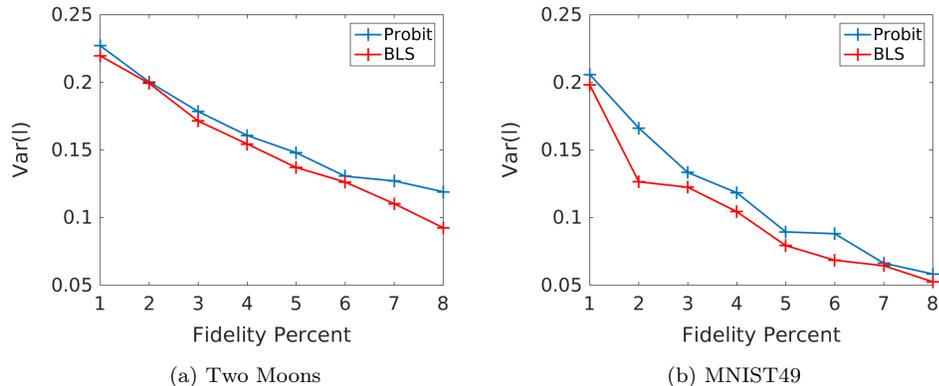


FIG. 7. Mean Posterior Variance as in (9) versus percentage of labelled points for the probit model and the BLS model applied to the Two Moons dataset and the 4-9 MNIST dataset. For two moons, we fix $N = 2,000$ and $\sigma = 0.06$. For each trial, we generate a realization of the two moons dataset while the MNIST dataset is fixed, and select at random a certain percentage of nodes as labelled. We run 20 trials for each percentage of fidelity, and average the mean posterior variance across trials. We set $\gamma = 0.1$ and $\beta = 0.1$ for both models.

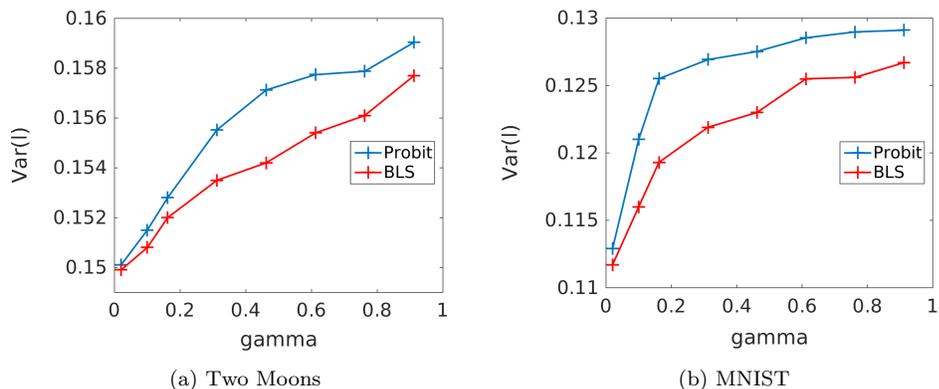


FIG. 8. Mean Posterior Variance as in (9) versus the noise parameter γ , applied to the Two Moons dataset and the 4-9 MNIST dataset. For two moons, we fix $N = 2,000$ and $\sigma = 0.06$. For each trial, we generate a realization of the two moons dataset while the MNIST dataset is fixed, and select randomly 4% percentage of nodes as labelled. We run 20 trials for each percentage of fidelity, and average the mean posterior variance across trials. We set γ as in the figure axis and $\beta = 0.3$ for both models.

not perform as well as spectral approximation: Algorithm 3 yields results close to the full posterior obtained from Algorithm 1. In contrast the spectral projection Algorithm 2 tends to underestimate the posterior variance, resulting in mean label values biased towards the values -1 or 1 .

5.3.2. Applications to Hyperspectral Imaging. In Figures 10 and 11 we apply the Bayesian level-set model to the hyperspectral image dataset; the results for probit are similar (when we use small γ) but have greater cost per step, because of the cdf evaluations required for probit. The figures show that the posterior mean

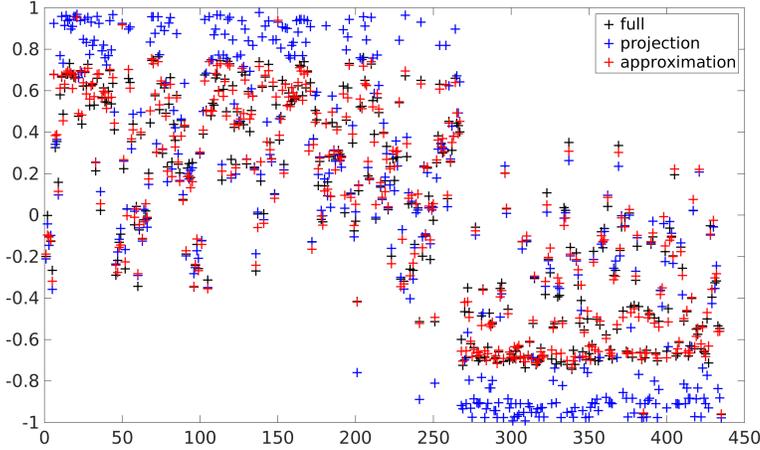


FIG. 9. Node-wise posterior mean s_j^l of the full Laplacian, spectral truncation, and spectral approximation method of the probit model with the voting records dataset. The horizontal axis denotes the index of the nodes (voters), and the vertical axis denotes the values of s_j^l of the node j . The per node mean absolute difference for spectral projection versus full is 0.1577; it is 0.0261 for spectral approximation versus full. We set $\gamma = 0.1$ and $\beta = 0.3$, and set the truncation level $\ell = 150$.

s_j^l is able to differentiate between different concentrations of the plume gas. We have also coloured pixels with $|s_j^l| < 0.4$ in red to highlight the regions with greater levels of uncertainty. We observe that the red pixels mainly lie in the edges of the gas plume, which conforms with human intuition. As in the voting records example in the previous subsection, the spectral approximation method has greater posterior uncertainty, demonstrated by the greater number of red pixels in Fig.10 compared to Fig.11. We conjecture that the spectral approximation is closer to what would be obtained by sampling the full distribution, but we have not verified this as the full problem is too large to readily sample.

In Figure 12 we study optimization via the MAP estimator of the Ginzburg-Landau model, employing the Algorithm 4. Assuming that the results of Bayesian level-set sampling are accurate classifiers we deduce that the Ginzburg-Landau MAP estimator, since similar, is also a good classifier. However it would be unfeasible as the basis for sampling on this problem, again because of multi-modality of the posterior.

5.4. Comparative Remarks About The Different Models. At a high level we have shown the following concerning the three models based on probit, level-set and Ginzburg-Landau:

- Probit and Bayesian level-set behave similarly, for posterior sampling, especially for small γ , since they formally coincide when $\gamma = 0$. Bayesian level set is considerably cheaper to implement in Matlab because the norm cdf evaluations required for probit are expensive; this property of probit could perhaps be addressed directly via dedicated programming.
- Probit and Bayesian level-set are superior to Ginzburg-Landau for posterior sampling; this is because probit has log-concave posterior, whilst Ginzburg-Landau is multi-modal.
- Ginzburg-Landau provides the best hard classifiers, when used as an opti-

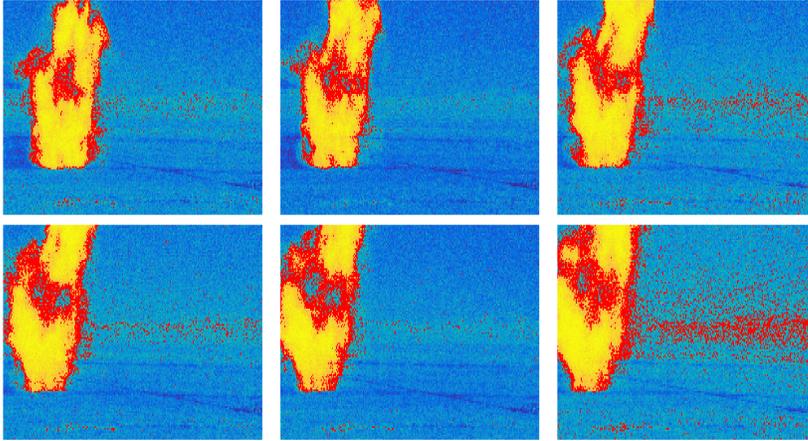


FIG. 10. Posterior mean s_j^l of the hyperspectral image dataset using Bayesian level-set model with **spectral approximation**. Each node is identified with the corresponding spatial pixel, and the values of s_j^l are plotted on a $[-1, 1]$ color scale on each pixel location. In addition, we highlight the regions of uncertain classification by coloring the pixels with $|s_j^l| < 0.4$ in red. The truncation level ℓ is set to be 40, and $\bar{\lambda} = 1.0$. We set $\gamma = 0.1$, $\beta = 0.08$ and use $M = 2 \times 10^4$ MCMC samples. We create the label data by subsampling 8,000 pixels ($\approx 2.8\%$ of the total) from the labelings obtained by spectral clustering.

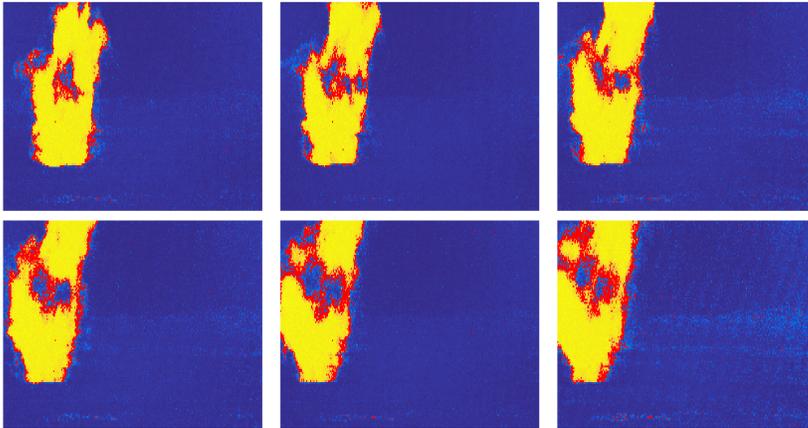


FIG. 11. Posterior mean s_j^l of the hyperspectral image dataset using Bayesian level-set model with **spectral projection**. Each node is identified with the corresponding spatial pixel, and the values of s_j^l are plotted on a $[-1, 1]$ color scale on each pixel location. In addition, we highlight the regions of uncertain classification by coloring the pixels with $|s_j^l| < 0.4$ in red. The truncation level ℓ is set to be 40. We set $\gamma = 0.1$, $\beta = 0.08$ and use $M = 2 \times 10^4$ MCMC samples. We create the label data by subsampling 8,000 pixels ($\approx 2.8\%$ of the total) from the labelings obtained by spectral clustering.

mizer (MAP estimator), and provided it is initialized well. However it behaves poorly when not initialized carefully because of multi-modal behaviour; in contrast probit has a convex objective function and hence a unique minimizer. (See supplementary materials for details of the relevant experiments.)

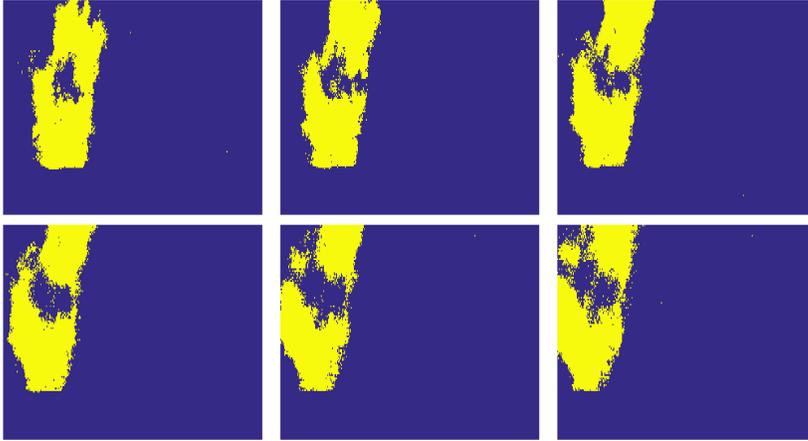


FIG. 12. Classification result of the hyperspectral image dataset using the Ginzburg-Landau MAP estimator. We identify each node with its corresponding spatial pixel, and plot the threshold value of the MAP estimator. We perform Algorithm 4 projected onto the first $N_{eig} = 40$ eigenvectors of the graph Laplacian. We set $\epsilon = 1$, $\gamma = 2.2$, and create the label data by subsampling 8,000 pixels ($\approx 2.8\%$ of the total) from the labelings obtained by spectral clustering.

Note that although probit has log concave posterior, and that this drives its computational benefits, it works on latent variable space, and not on label space where the Ginzburg-Landau method operates. Although we have not reported results using the atomic noise model, these are similar to Bayesian level-set due to the structurally identical form of the log likelihood, which is piecewise constant for both cases.

We expand on the details of these conclusions by studying run times of the algorithms. All experiments are done on a 1.5GHz machine with Intel Core i7. In Table 2, we compare the running time of the MCMC for different models on various datasets. We use an a posteriori condition on the samples $u^{(k)}$ to empirically determine the sample size M needed for the MCMC to converge. Note that this condition is by no means a replacement for a rigorous analysis of convergence using auto-correlation, but is designed to provide a ballpark estimate of the speed of these algorithms on real applications. We now define the a posteriori condition used. Let the approximate samples be $\{u^{(k)}\}$. We define the cumulative average as $\tilde{u}^{(k)} = \frac{1}{k} \sum_{j=1}^k u^{(j)}$, and find the first k such that

$$(17) \quad \|\tilde{u}^{(kT)} - \tilde{u}^{((k-1)T)}\| \leq \mathbf{tol},$$

where \mathbf{tol} is the tolerance and T is the number of iterations skipped. We set $T = 5000$, and also tune the stepsize parameter β such that the average acceptance probability of the MCMC is over 50%. We choose the model parameters according to the experiments in the sections above so that the posterior mean gives a reasonable classification result.

We note that the number of iterations needed for the Ginzburg-Landau model is much higher compared to Probit and the Bayesian level-set (BLS) method; this is caused by the presence of multiple local minima in Ginzburg-Landau, in contrast to

⁵According to the reporting in [30].

Data (Tol)	Voting Records $\text{tol} = 1 \times 10^{-3}$	MNIST49 $\text{tol} = 1.5 \times 10^{-3}$	Hyperspectral $\text{tol} = 2 \times 10^{-2}$
(N)	$N = 435$	$N \approx 1.1 \times 10^4$	$N \approx 2.9 \times 10^5$
(Neig)	$Neig = 435$	$Neig = 300$	$Neig = 50$
(J)	$J = 5$	$J = 440$	$J = 8000$
Preprocessing	$t = 0.7s$	$t = 50.8s$	$t < 60s^5$
probit	$t = 8.9s, M = 10^4$	$t = 176.4s, M = 1.5 \times 10^4$	$5410.3s, M = 1.5 \times 10^4$
BLS	$t = 2.7s, M = 10^4$	$t = 149.1s, M = 1.5 \times 10^4$	$970.8s, M = 1.5 \times 10^4$
GL	$t = 161.4s, M = 1.8 \times 10^5$	-	-

TABLE 2

Timing for MCMC methods. We report both the number of samples M and the running time of the algorithm t . The time for GL on MNIST and Hyperspectral is omitted due to running time being too slow. J denotes the number of fidelity points used. For the voting records, we set $\gamma = 0.2, \beta = 0.4$ for Probit and BLS, and $\gamma = 1, \beta = 0.1$ for Ginzburg-Landau. For MNIST, we set $\gamma = 0.1, \beta = 0.4$. For Hyperspectral, we set $\gamma = 1.0$, and $\beta = 0.1$.

the log concavity of probit. Probit is slower than BLS due to the fact that evaluations of the CDF function for Gaussians is slow.

6. Conclusions and Future Directions. We have introduced a Bayesian approach to uncertainty quantification for graph-based classification methods. We develop algorithms to sample the posterior and to compute MAP estimators and, through numerical experiments on a suite of applications, we have investigated the properties of the different Bayesian models, and the algorithms explored to study them.

Some future directions of this work include improvement of the current inference method, connections between the different models in this paper, and generalization to multiclass classification. For example, one could accelerate the current scheme by applying gradient information in the MCMC proposal while maintaining the dimension independence of pCN. Alternatively, we could apply techniques other than MCMC such as variational methods to approximate the posterior expectation. The current MCMC method is also slow for the Ginzburg-Landau model due to the presence of local extrema. An interesting question is whether there are more efficient means of sampling from this distribution. One could study the small noise limit of the various posterior measures from different models. For example, the small noise limit of the probit and Bayesian level-set models coincide, and hence one would expect the two models to have similar posterior distributions when γ is small. For multiclass classification, one could vectorize the latent variable (as in existing non-Bayesian multiclass methods [18, 29]), and applying multi-dimensional analogues of the likelihood functions used in this paper. Hierarchical methods could also be applied to account for the uncertainty in the various hyperparameters such as the label noise γ , or the length scale ϵ in the Ginzburg-Landau model. Finally, we could study in more detail the effects of either the spectral projection or the approximation method. One could attempt to quantify the quality of the spectral approximation/truncation algorithm in terms of the posterior distribution, either analytically on some tractable toy examples, or empirically on a suite of representative problems.

Acknowledgements AMS is grateful to Omiros Papaspiliopoulos for illuminating discussions about probit and the atomic noise model.

- [1] J. H. ALBERT AND S. CHIB, *Bayesian analysis of binary and polychotomous response data*, Journal of the American Statistical Association, 88 (1993), pp. 669–679.
- [2] M. BELKIN, I. MATVEEVA, AND P. NIYOGI, *Regularization and semi-supervised learning on large graphs*, in International Conference on Computational Learning Theory, Springer, 2004, pp. 624–638.
- [3] 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 (2006), pp. 2399–2434.
- [4] M. BERTHOD, Z. KATO, S. YU, AND J. ZERUBIA, *Bayesian image classification using Markov random fields*, Image and Vision Computing, 14 (1996), pp. 285–295.
- [5] A. L. BERTOZZI AND A. FLENNER, *Diffuse interface models on graphs for classification of high dimensional data*, Multiscale Modeling & Simulation, 10 (2012), pp. 1090–1118.
- [6] A. BESKOS, G. ROBERTS, AND A. STUART, *Optimal scalings for local metropolis-hastings chains on nonproduct targets in high dimensions*, The Annals of Applied Probability, (2009), pp. 863–898.
- [7] A. BESKOS, G. ROBERTS, A. M. STUART, AND J. VOSS, *MCMC methods for diffusion bridges.*, Stochastics and Dynamics, 8 (2008), pp. 319–350.
- [8] A. BLUM AND S. CHAWLA, *Learning from labeled and unlabeled data using graph mincuts*, (2001).
- [9] Y. BOYKOV, O. VEKSLER, AND R. ZABIH, *Markov random fields with efficient approximations*, in Computer vision and pattern recognition, 1998. Proceedings. 1998 IEEE computer society conference on, IEEE, 1998, pp. 648–655.
- [10] Y. BOYKOV, O. VEKSLER, AND R. ZABIH, *Fast approximate energy minimization via graph cuts*, IEEE Transactions on pattern analysis and machine intelligence, 23 (2001), pp. 1222–1239.
- [11] Y. Y. BOYKOV AND M.-P. JOLLY, *Interactive graph cuts for optimal boundary & region segmentation of objects in nd images*, in Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on, vol. 1, IEEE, 2001, pp. 105–112.
- [12] J. B. BROADWATER, D. LIMSUI, AND A. K. CARR, *A primer for chemical plume detection using LWIR sensors*, Technical Paper, National Security Technology Department, Las Vegas, NV, (2011).
- [13] T. BÜHLER AND M. HEIN, *Spectral clustering based on the graph p-laplacian*, in Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 81–88.
- [14] F. R. CHUNG, *Spectral graph theory*, vol. 92, American Mathematical Soc., 1997.
- [15] 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 (2013), pp. 424–446.
- [16] E. DAHLHAUS, D. S. JOHNSON, C. H. PAPADIMITRIOU, P. D. SEYMOUR, AND M. YANNAKAKIS, *The complexity of multiway cuts*, in Proceedings of the twenty-fourth annual ACM symposium on theory of computing, ACM, 1992, pp. 241–251.
- [17] C. FOWLKES, S. BELONGIE, F. CHUNG, AND J. MALIK, *Spectral grouping using the Nyström method*, IEEE transactions on pattern analysis and machine intelligence, 26 (2004), pp. 214–225.
- [18] C. GARCIA-CARDONA, E. MERKURJEV, A. L. BERTOZZI, A. FLENNER, AND A. G. PERCUS, *Multiclass data segmentation using diffuse interface methods on graphs*, IEEE transactions on pattern analysis and machine intelligence, 36 (2014), pp. 1600–1613.
- [19] D. K. HAMMOND, P. VANDERGHEYNST, AND R. GRIBONVAL, *Wavelets on graphs via spectral graph theory*, Applied and Computational Harmonic Analysis, 30 (2011), pp. 129–150.
- [20] J. HARTOG AND H. VAN ZANTEN, *Nonparametric bayesian label prediction on a graph*, arXiv preprint arXiv:1612.01930, (2016).
- [21] D. J. HIGHAM AND M. KIBBLE, *A unified view of spectral clustering*, University of Strathclyde mathematics research report, 2 (2004).
- [22] C. C. HOLMES, L. HELD, ET AL., *Bayesian auxiliary variable models for binary and multinomial regression*, Bayesian Analysis, 1 (2006), pp. 145–168.
- [23] H. HU, J. SUNU, AND A. L. BERTOZZI, *Multi-class graph mumford-shah model for plume detection using the MBO scheme*, in Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer, 2015, pp. 209–222.
- [24] M. A. IGLESIAS, Y. LU, AND A. M. STUART, *A Bayesian Level Set Method for Geometric Inverse Problems*, Interfaces and Free Boundary Problems, arXiv preprint arXiv:1504.00313, (2015).
- [25] A. KAPOOR, Y. QI, H. AHN, AND R. PICARD, *Hyperparameter and kernel learning for graph based semi-supervised classification*, in NIPS, 2005, pp. 627–634.
- [26] Y. LECUN, C. CORTES, AND C. J. BURGES, *The MNIST database of handwritten digits, online at <http://yann.lecun.com/exdb/mnist/>*, 1998.

- [27] S. Z. LI, *Markov random field modeling in computer vision*, Springer Science & Business Media, 2012.
- [28] A. MADRY, *Fast approximation algorithms for cut-based problems in undirected graphs*, in Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on, IEEE, 2010, pp. 245–254.
- [29] E. MERKURJEV, T. KOSTIC, AND A. L. BERTOZZI, *An MBO scheme on graphs for classification and image processing*, SIAM Journal on Imaging Sciences, 6 (2013), pp. 1903–1930.
- [30] E. MERKURJEV, J. SUNU, AND A. L. BERTOZZI, *Graph mbo method for multiclass segmentation of hyperspectral stand-off detection video*, in Image Processing (ICIP), 2014 IEEE International Conference on, IEEE, 2014, pp. 689–693.
- [31] R. NEAL, *Regression and classification using gaussian process priors*, Bayesian Statistics, 6, p. 475. Available at <http://www.cs.toronto.edu/radford/valencia.abstract.html>.
- [32] H. OWHADI, C. SCOVEL, AND T. SULLIVAN, *On the brittleness of bayesian inference*, SIAM Review, 57 (2015), pp. 566–582.
- [33] H. OWHADI, C. SCOVEL, T. J. SULLIVAN, M. MCKERNS, AND M. ORTIZ, *Optimal uncertainty quantification*, SIAM Review, 55 (2013), pp. 271–345.
- [34] O. PAPASPILIOPOULOS, *Private communication*, 2016.
- [35] G. O. ROBERTS, A. GELMAN, W. R. GILKS, ET AL., *Weak convergence and optimal scaling of random walk metropolis algorithms*, The Annals of Applied Probability, 7 (1997), pp. 110–120.
- [36] D. I. SHUMAN, M. FARAJI, AND P. VANDERGHEYNST, *Semi-supervised learning with spectral graph wavelets*, in Proceedings of the International Conference on Sampling Theory and Applications (SampTA), no. EPFL-CONF-164765, 2011.
- [37] R. C. SMITH, *Uncertainty quantification: theory, implementation, and applications*, vol. 12, SIAM, 2013.
- [38] T. J. SULLIVAN, *Introduction to uncertainty quantification*, vol. 63, Springer, 2015.
- [39] N. G. TRILLOS AND D. SLEPČEV, *A variational approach to the consistency of spectral clustering*, Applied and Computational Harmonic Analysis, (2016).
- [40] Y. VAN GENNIP AND A. L. BERTOZZI, *Γ -convergence of graph Ginzburg-Landau functionals*, Advances in Differential Equations, 17 (2012), pp. 1115–1180.
- [41] U. VON LUXBURG, *A tutorial on spectral clustering*, Statistics and Computing, 17 (2007), pp. 395–416.
- [42] U. VON LUXBURG, M. BELKIN, AND O. BOUSQUET, *Consistency of spectral clustering*, The Annals of Statistics, (2008), pp. 555–586.
- [43] C. K. WILLIAMS AND C. E. RASMUSSEN, *Gaussian Processes for Regression*, (1996).
- [44] C. K. WILLIAMS AND M. SEEGER, *Using the Nyström method to speed up kernel machines*, in Proceedings of the 13th International Conference on Neural Information Processing Systems, MIT press, 2000, pp. 661–667.
- [45] D. XIU, *Numerical Methods For Stochastic Computations: A Spectral Method Approach*, Princeton University Press, 2010.
- [46] L. ZELNIK-MANOR AND P. PERONA, *Self-tuning spectral clustering*, in Advances in neural information processing systems, 2004, pp. 1601–1608.
- [47] D. ZHOU, O. BOUSQUET, T. N. LAL, J. WESTON, AND B. SCHÖLKOPF, *Learning with local and global consistency*, Advances in neural information processing systems, 16 (2004), pp. 321–328.
- [48] X. ZHU, *Semi-supervised learning literature survey*, Technical Report TR1530.
- [49] X. ZHU, Z. GHARAMANI, J. LAFFERTY, ET AL., *Semi-supervised learning using Gaussian fields and harmonic functions*, in ICML, vol. 3, 2003, pp. 912–919.
- [50] X. ZHU, J. D. LAFFERTY, AND Z. GHARAMANI, *Semi-supervised learning: From Gaussian fields to Gaussian processes*, (2003).

Appendix.

6.1. Spectral Properties of L . The spectral properties of L are relevant to the spectral projection and approximation algorithms from the previous section. Figure 2 shows the spectra for our four examples. Note that in all cases the spectrum is contained in the interval $[0, 2]$, consistent with the theoretical result in [14, Lemma 1.7, Chapter 1]. The size of the eigenvalues near to 0 will determine the accuracy of the spectral projection algorithm. The rate at which the spectrum accumulates at a value near 1, an accumulation which happens for all but the MNIST data set in our four examples, affects the accuracy of the spectral approximation algorithm. There is theory that goes some way towards justifying the observed accumulation; see [42, Proposition 9, item 4]. This theory works under the assumption that the features x_j are i.i.d samples from some fixed distribution, and the graph Laplacian is constructed from weights $w_{ij} = k(x_i, x_j)$, and k satisfies symmetry, continuity and uniform positivity. As a consequence the theory does not apply to the graph construction used for the MNIST dataset since the K -nearest neighbor graph is local; empirically we find that this results in a graph violating the positivity assumption on the weights. This explains why the MNIST example does not have a spectrum which accumulates at a value near 1. In the case where the spectrum does accumulate at a value near 1, the rate can be controlled by adjusting the parameter τ appearing in the weight calculations; in the limit $\tau = \infty$ the graph becomes an unweighted complete graph and its spectrum comprises the the two points $\{0, \lambda\}$ where $\lambda \rightarrow 1$ as $n \rightarrow \infty$ – see Lemma 1.7 in Chapter 1 of [14].

6.2. MAP Estimation as Semi-supervised Classification Method. We first prove the convexity of the probit negative log likelihood.

PROPOSITION 1. *Let $J_p(u)$ be the MAP estimation function for the probit model:*

$$J_p(u) = \frac{1}{2} \langle u, Pu \rangle - \sum_{j \in Z'} \log \left(\Psi(y(j)u(j); \gamma) \right).$$

If $y(j) \in \{\pm 1\}$ for all j then J_p is a convex function in the variable u .

Proof. Since P is semi positive definite, it suffices to show that

$$\sum_{j \in Z'} \log \left(\Psi(y(j)u(j); \gamma) \right)$$

is convex. Thus, since $y(j) \in \{\pm 1\}$ for all j , it suffices to show that $\log \left(\Psi(x; \gamma) \right)$ is concave with respect to x . Since

$$\Psi(x; \gamma) = \frac{1}{\sqrt{2\pi\gamma}} \int_{-\infty}^x \exp\left(\frac{-t^2}{2\gamma^2}\right) dt,$$

we have $\Psi(\gamma x; \gamma) = \Psi(x; 1)$. Since scaling x by a constant doesn't change convexity, it suffices to consider the case $\gamma = 1$. Taking the second derivative with $\gamma = 1$, we see that it suffices to prove that, for all $x \in \mathbb{R}$ and all $\gamma > 0$,

$$(18) \quad \Psi''(x; 1)\Psi(x; 1) - \Psi'(x; 1)\Psi'(x; 1) < 0.$$

Plugging in the definition of Ψ , we have

$$(19) \quad \Psi''(x; 1)\Psi(x; 1) - \Psi'(x; 1)\Psi'(x; 1) = \frac{-1}{2\pi} \exp\left(\frac{-x^2}{2}\right) \left(x \int_{-\infty}^x \exp\left(\frac{-t^2}{2}\right) dt + \exp\left(\frac{-x^2}{2}\right) \right).$$

Clearly the expression in equation (19) is less than 0 for $x \geq 0$. For the case $x < 0$, divide equation (19) by $\frac{1}{2\pi} \exp(\frac{-x^2}{2})$ and note that this gives

$$(20) \quad -x \int_{-\infty}^x \exp(\frac{-t^2}{2}) dt - \exp(\frac{-x^2}{2}) = -x \int_{-\infty}^x \exp(\frac{-t^2}{2}) dt + \int_{-\infty}^x t \exp(\frac{-t^2}{2}) dt \\ = \int_{-\infty}^x (t - x) \exp(\frac{-t^2}{2}) dt < 0$$

and the proof is complete. \square

The probit MAP estimator thus has a considerable computational advantage over the Ginzburg-Landau MAP estimator, because the latter is not convex and, indeed, can have large numbers of minimizers. We now discuss numerical results designed to probe the consequences of convexity, or lack of it, for classification accuracy. The purpose of these experiments is not to match state-of-art results for classification, but rather to study properties of the MAP estimator when varying the feature noise and the percentage of labelled data.

We employ the two moons and the MNIST (4, 9) data sets. The methods are evaluated on a range of values for the percentage of labelled data points, and also for a range of values of the feature variance σ in the two moons dataset. The experiments are conducted for 100 trials with different initializations (both two moons and MNIST (4, 9)) and different data realizations (for two moons only). In Figure 13, we plot the median classification accuracy with error bars from the 100 trials against the feature variance σ for the two moons dataset. As well as Ginzburg-Landau and probit classification, we also display results from spectral clustering based on thresholding the Fiedler eigenvector. The percentage of fidelity points used is 0.5%, 1%, and 3% for each column. We do the same in Figure 14 for the 4-9 MNIST data set against the same percentages of labelled points.

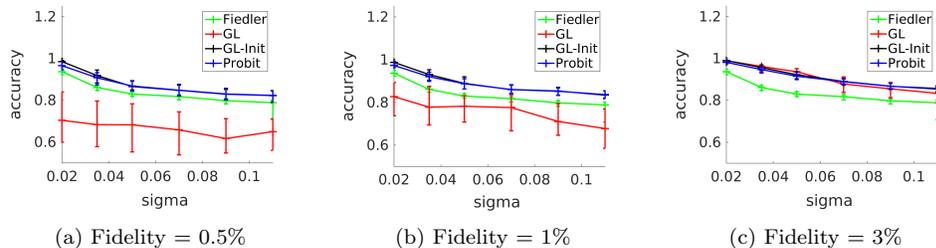


FIG. 13. Classification accuracy of different algorithms for Two Moons Dataset compared with σ and percentage of labelled nodes, with $N = 2,000$. The algorithms used are: Ginzburg-Landau MAP estimator with random initialization, Ginzburg-Landau with initialization given by probit model, probit MAP estimation, and spectral clustering (thresholding the Fiedler vector). For each trial, we generate a realization of the two moons dataset with given σ and select randomly a certain percentage of nodes as fidelity, and a total of 50 trials are run for each combination of parameters. We use spectral projection with number of eigenvectors $N_{eig} = 150$. We plot the median accuracy along with error bars indicating the 25 and 75-th quantile of the classification accuracy of each method. We set $\gamma = 0.1$ for the probit model, and $\gamma = 1.0$, $\epsilon = 1.0$ for Ginzburg-Landau.

The non-convexity of the Ginzburg-Landau model can result in large variance in classification accuracy; the extent of this depends on the percentage of observed labels. The existence of sub-optimal local extrema causes the large variance. If

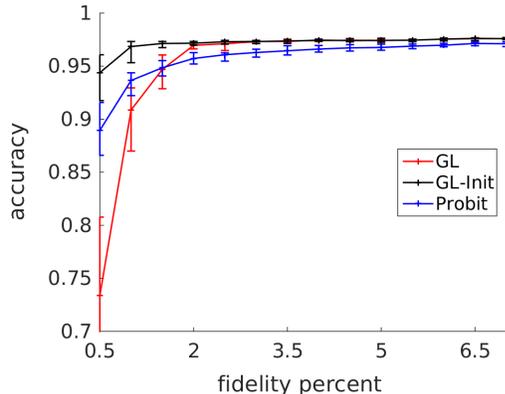


FIG. 14. Classification accuracy of different algorithms for the 4-9 MNIST dataset versus percentage of labelled nodes. The algorithms used are: Ginzburg-Landau with random initialization, Ginzburg-Landau with initialization given by probit model, probit MAP estimation. For each trial, we select randomly a certain percentage of nodes as fidelity, and a total of 50 trials are run. We use spectral projection with number of eigenvectors $N_{\text{eig}} = 300$. We plot the median accuracy along with error bars indicating the 25 and 75-th quantile of the classification accuracy of each method. We set $\gamma = 0.1$ for the probit model, and $\gamma = 1.0$, $\epsilon = 1.0$ for Ginzburg-Landau.

initialized without information about the classification, Ginzburg-Landau can perform very badly in comparison with probit. On the other hand we find that the best performance of the Ginzburg-Landau model, when initialized at the probit minimizer, is typically slightly better than the probit model.

We note that the probit model is convex and theoretically should have results independent of the initialization. However, we see there are still small variations in the classification result from different initializations. This is due to slow convergence of gradient methods caused by the flat-bottomed well of the probit log-likelihood. As mentioned above this can be understood by noting that, for small gamma, probit and level-set are closely related and that the level-set MAP estimator does not exist – minimizing sequences converge to zero, but the infimum is not attained at zero.