Journal of Machine Learning Research 21 (2020) 1-55 Submitted 6/19; Revised 3/20; Published 7/20

Consistency of Semi-Supervised Learning Algorithms on Graphs: Probit and One-Hot Methods

Franca Hoffmann FKOH@CALTECH.EDU
Bamdad Hosseini BAMDADH@CALTECH.EDU
Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125, USA

Zhi Ren ZREN@MIT.EDU
Department of Mathematics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Andrew M. Stuart ASTUART@CALTECH.EDU
Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125, USA

Editor: Bert Huang

Abstract

Graph-based semi-supervised learning is the problem of propagating labels from a small number of labelled data points to a larger set of unlabelled data. This paper is concerned with the consistency of optimization-based techniques for such problems, in the limit where the labels have small noise and the underlying unlabelled data is well clustered. We study graph-based probit for binary classification, and a natural generalization of this method to multi-class classification using one-hot encoding. The resulting objective function to be optimized comprises the sum of a quadratic form defined through a rational function of the graph Laplacian, involving only the unlabelled data, and a fidelity term involving only the labelled data. The consistency analysis sheds light on the choice of the rational function defining the optimization.

Keywords: Semi-supervised learning, classification, consistency, graph Laplacian, probit, spectral analysis.

1. Introduction

Semi-supervised learning (SSL) is the problem of labelling all the points in a data set, by leveraging correlations and geometric information in the data points, together with explicit knowledge of a subset of noisily observed labels. The primary goal of this article is to analyze the probit and one-hot methods for transductive SSL. We elaborate conditions under which these methods consistently recover the correct labels attached to the unlabelled data set. We do this in an idealized setting in which the unlabelled data is approximately clustered, and there is an unobserved latent variable which determines labels and which is observed in a small noise regime. We prove consistency in a limit in which the data becomes more

clustered and the label noise goes to zero. The formulation and analysis demonstrates how ideas from unsupervised learning and, in particular spectral clustering, can be used as prior information; this prior information is enhanced, or sets up a competition with, labelled data. In so doing, our analysis also elucidates the role of parameter choices made when setting up the balance between labelled and unlabelled data. Furthermore, we exhibit useful properties of the probit and one-hot methods, including a representer theorem for the classifier, and a natural dimension reduction which follows from this theorem and is impactful in practice.

1.1. Background and Literature Review

We start by giving informal statements of the problem to be solved, and a brief literature review. Consider a set of nodes $Z = \{1, \cdots, N\}$ and an associated set of *feature vectors* $X = \{x_1, x_2, \cdots, x_N\}$. Each feature vector x_j is assumed to be a point in $\mathbb{R}^d$. X may thus be viewed as a function $X : Z \mapsto \mathbb{R}^d$ or as an element of $\mathbb{R}^{d\times N}$. We refer to X as *unlabelled data*. Suppose there exists a function $l : Z \mapsto \{1, 2, \cdots, M\}$ that assigns one of M distinct labels to each point in Z. That is, for every point $j \in Z$ the value $l(j) = m$ indicates that j belongs to class m or is *labelled* as m. Throughout this article we assume that every point in Z belongs to one class only.

Now let $Z' \subseteq Z$ be a subset of the nodes with $|Z'| = J \leq N$ and define $y : Z' \mapsto \{1, 2, \cdots, M\}$ to be a *noisily observed label* of each point in Z'. We refer to y as *labelled data*. With this setup we may define the SSL problem.

Problem 1 (Semi-Supervised Learning) *Suppose Z, Z', X and y are known. Find $l : Z \mapsto \{1, 2, \cdots, M\}$.* ◇

In order to solve this problem, which is highly ill-posed, it is necessary to introduce some form of regularity on the labels, guided by the correlations in X for example, and to make assumptions about the errors in the labels provided. One approach, which we study here, is to assume that the labels on Z are defined through a latent variable $u : Z \mapsto \mathbb{R}^M$, whose regularity is defined through the unlabelled data X, and a function $S : \mathbb{R}^M \mapsto \{1, 2, \cdots, M\}$. Specifically we assume that there is a ground truth function $u^\dagger : Z \mapsto \mathbb{R}^M$ for which

$$y(j) := S\big(u^\dagger(j) + \eta(j)\big), \qquad j \in Z', \tag{1}$$

where $\eta(j) \overset{iid}{\sim} \psi$ and ψ is the Lebesgue density of a zero-mean random variable on $\mathbb{R}^M$. We may now introduce the following relaxation of the SSL problem.

Problem 2 (Relaxed Semi-Supervised Learning) *Suppose Z, Z', X and y are known, together with the function S and the density ψ. Find $u : Z \mapsto \mathbb{R}^M$ and define $l = S \circ u : Z \mapsto \{1, 2, \cdots, M\}$.* ◇

In Problem 3 below we will define a class of optimization functionals for u, giving an explicit instantiation of Problem 2, and focus on the resulting optimization problems in our analysis. Before doing so we give a literature review explaining the context for this optimization approach.

The consistency of classification methods in the setting of supervised learning is well-developed; see Tewari and Bartlett (2007) for a literature review and results applying to both

binary and multi-class classification, as well as the preceding work in Steinwart (2001, 2005); Wu and Zhou (2006) which establishes the problem in the framework of Vapnik (1995). The paper of Xu et al. (2009) discusses the robustness of such supervised classification methods, allowing for a small fraction of adversarially labelled data points. There has been some recent analysis of logistic regression, and the reader may access the literature on this subject via the recent papers of Candès and Sur (2020); Sur and Candès (2019). All of this work on supervised classification focuses on the large data/large number of features setting, and often starts from assumptions that the unlabelled data is linearly separated. None of it leverages the power of graph-based techniques to extract geometric information in large unlabelled data sets. To make the connection to graph-based techniques we need to discuss unsupervised graph-based learning (Belkin and Niyogi, 2001; Von Luxburg, 2007). This is a subject that has seen significant analysis in relation to consistency. Spielman and Teng (1996, 2007) perform a careful analysis of the spectral gaps of graph Laplacians resulting from clustered data, studying recursive methods for multi-class clustering. Ng et al. (2001) introduced a way of thinking about, and analyzing, multi-class unsupervised learning based on perturbing a perfectly clustered case; we will leverage similar ideas in our work on SSL. Von Luxburg et al. (2008) introduced the idea of studying the consistency of spectral clustering in the limit of large i.i.d. data sets in which the graph Laplacian converges to a limiting integral operator. García-Trillos and Slepčev (2018); García-Trillos et al. (2019) have taken this further by working with localizing weight functions designed so that the limit of the graph Laplacian is a differential operator.

SSL is a methodology which combines the methods of unsupervised learning and of supervised classification. According to the definition in Kostopoulos et al. (2018) "SSL can be categorized into two somewhat different settings, namely inductive and transductive learning ... inductive SSL attempts to predict the labels on unseen future data, while transductive SSL attempts to predict the labels on unlabeled instances taken from the training set." In this paper our focus is on transductive SSL. Initial attempts to solve the SSL problem employed combinatorial algorithms. Based on an explicit mathematical formulation stemming from Problem 1; see Blum and Chawla (2001). Later Zhu et al. (2003a,b) introduced a relaxation similar to Problem 2. Their approach is most easily described in the binary case in which they assume $S : \mathbb{R} \mapsto \mathbb{R}$ is the identity function and the labels are given in the form ± 1. From a modeling viewpoint this approach is unnatural because the categorical data is assumed to also lie in the real-valued space of the latent variable. Bertozzi and Flenner (2012) introduced an interesting relaxation of this assumption, by means of a Ginzburg-Landau penalty term which favours real-values close to ± 1 but does not enforce the categorical values ± 1 exactly. The probit approach to classification, described in the classic text of Rasmussen and Williams (2006) on Gaussian process regression, does not make the unnatural modeling assumption underlying Zhu's work; instead it is based on taking S to be the sign function. However the basic form of probit in Rasmussen and Williams (2006) does not use unlabelled data to extend labels outside the labelled data set, but instead does so through regular Gaussian process regression: inductive SSL.

The extension of the probit method to graph-based transductive SSL is described in Bertozzi et al. (2018), where both Bayesian and optimization-based formulations are de-

scribed; in that paper, (1) is also generalized to the level set form

$$y(j) := S\big(u^\dagger(j)\big) + \eta(j), \qquad j \in Z', \tag{2}$$

and a Bayesian formulation of the Ginzburg-Landau relaxation of Bertozzi and Flenner (2012) is introduced. The close relationship between level set and probit formulations is discussed in Dunlop et al. (2020). Belkin et al. (2004); Belkin and Niyogi (2001); Belkin et al. (2006) demonstrate how both Gaussian process regression and graph-based SSL can be used simultaneously.All of the approaches which followed from the work of Zhu are readily generalized from the binary case to the multi-class setting, using the idea of one-hot encoding, explained in detail in Subsection 3.1, in which each label is identified with a standard unit basis vector in $\mathbb{R}^M$.

A large number of approaches to SSL have been developed in the literature and a detailed discussion of all of them is outside the scope of this article. We refer the reader to the review articles of Zhu (2005) and Kostopoulos et al. (2018) for, respectively, the state-of-the-art in 2005 and a more recent appraisal of the field that categorizes various inductive and transductive approaches to SSL and semi-supervised regression. The idea of regularization by graph Laplacians for SSL was developed in different contexts such as manifold regularization (Belkin et al., 2006), Tikhonov regularization (Belkin et al., 2004) and local learning regularization (Wu and Schölkopf, 2007). However, while graph regularization methods are widely applied in practice the rigorous analysis of their properties, and in particular asymptotic consistency, is not well-developed within the context of SSL. Indeed, to the best of our knowledge the consistency analysis of the probit and one-hot methods has not been tackled before. SSL may be viewed as a method for boosting, refining or questioning unsupervised graph-based learning, through labelling information; our analysis sheds light on this process.

There has been other analysis of SSL methods, not concerning consistency. Dunlop et al. (2020) studied the large data and zero noise limits of the probit method. They derived a continuum inverse problem using the methodology of García-Trillos et al. (2019); García-Trillos and Slepčev (2018) that characterizes SSL when the number of vertices of the graph and the number of observed labels is fixed, or goes to infinity in a manner insuring a fixed fraction of labels. The authors also studied the zero noise limit of probit and level-set methods for SSL and showed that both problems approach the same limit as the noise variance goes to zero.

1.2. Problem Setup and Preliminaries

Our focus in this paper is on the analysis of algorithms built from the introduction of real-(vector)-valued latent functions, leading to precise mathematical formulations of Problem 2. To make actionable algorithms we need to specify precisely how the unlabelled data X and the labelled data y are used. The approach we study here is to define the desired latent variable u as the minimizer of a function comprised of two terms, one of which enforces correlations and geometric information in the unlabelled data X, and the other which enforces consistency with the label data y, on the assumption that they are related to u as in (1). To this end we view X as a point cloud in $\mathbb{R}^d$ and associate a weight matrix $W = (w_{ij})$ to tuples (x_i, x_j) in $X \times X$. The weights w_{ij}, which are assumed to be non-negative, are chosen

to measure affinities between x_i and x_j. Since similarity between data points is a symmetric relationship, we assume $w_{ij} = w_{ji}$ so that W is a symmetric matrix and define a proximity graph $G = \{X, W\}$ with vertices X and edge weights W. From W we will construct a covariance operator C on spaces of functions $H = \{u : Z \mapsto \mathbb{R}^M\}$, using a graph Laplacian implied by W. We also define a misfit function $\Phi(\cdot;\cdot) : H \times \{1, \cdots, M\}^J \mapsto \mathbb{R}$ which encodes the assumption (1) about the relationship between the labels and the latent function. With these objects we then formulate the SSL problem as a regularized optimization problem.

Problem 3 (Relaxed Semi-Supervised Learning as Optimization) *Suppose Z, Z', X and y are known, together with the function S, the covariance operator C and the misfit Φ. Find the function u^* defined by*

$$u^* = \arg\min_{u \in H} \frac{1}{2}\langle u, C^{-1}u\rangle_H + \Phi(u; y). \tag{3}$$

$\diamond$

This optimization problem may be viewed as the MAP estimator associated to the Bayesian inverse problem of finding the distribution of $u|y$ when the prior on u is a Gaussian random measure on H with covariance C and $\Phi(u; y)$ is the negative log-likelihood of y conditioned on u, i.e.

$$\mathbb{P}(y|u) \propto \exp\big(-\Phi(u; y)\big), \qquad \text{assuming} \qquad y(j) = S\big(u(j) + \eta(j)\big). \tag{4}$$

We refer to Φ as the *likelihood potential*.

1.3. Main Contributions

The key question at the heart of this article is to identify conditions under which the minimizer u^* of Problem 3 correctly identifies the labels. To this end, we define the following notion of consistency.

Definition 1 (SSL Asymptotic Consistency) *We say that Problem 3 is asymptotically consistent if, for all $j \in Z$,*

$$S(u^*(j)) \xrightarrow{\text{a.s.}} S(u^\dagger(j)), \qquad \text{as} \qquad \text{std}(\eta(j)) \downarrow 0,$$

where $u^\dagger$ is the latent variable underlying the labelled data (1).

In the above and throughout the rest of the article $\xrightarrow{\text{a.s.}}$ denotes almost sure (a.s.) convergence with respect to a common probability space on which the measurement noise $\eta(j)$ are defined (see Subsections 2.6 and 3.5 for a formal discussion of this mode of convergence). We primarily focus on the probit and one-hot methods for SSL, corresponding to specific choices of the function S. As mentioned earlier probit is an optimization approach for binary classification that formulates Problem 3 with $M = 2$. The one-hot method is a generalization of probit for multi-class classification when $M \geq 2$. We outline these methods in detail in Sections 2 and 3. We show that probit and one-hot methods are asymptotically consistent in the case where the graph G is nearly-disconnected in the following sense.

Definition 2 (Nearly-Disconnected Graph) *A weighted graph $G = \{X, W\}$ is nearly-disconnected with K clusters if there exist connected components $\tilde{G}_k = \{\tilde{X}_k, \tilde{W}_k\}$ for $k \in \{1, \cdots, K\}$ so that the edges within each $\tilde{G}_k$ are $\mathcal{O}(1)$, but the edges between elements in different $\tilde{G}_k$ are $\mathcal{O}(\epsilon)$ for a small parameter $\epsilon > 0$. In other words, up to a reordering of the index set Z, the matrix W is nearly block diagonal.*

Working in such a setting is a natural way of representing nearly clustered data, and was exploited by Ng et al. (2001) concerning unsupervised learning. The number of clusters K is an inherent geometric property of the unlabelled data X; determining a suitable choice of K in practice can be challenging and depends on the scale one is interested in. In the following informal statement of our main result we assume that each component $\tilde{G}_k$ is associated with at least one pre-assigned label. The result shows that if G is nearly-disconnected and the ground truth function $u^\dagger$ assigns the same label to all points within each component $\tilde{G}_k$ then the probit and one-hot methods are asymptotically consistent for an appropriate choice of matrix C so long as at least one label is observed in each component $\tilde{G}_k$. Below, L denotes the *graph Laplacian*, a discrete diffusion operator acting on functions defined on the graph G, see Section 2.3 for a precise definition.

Theorem 3 (Consistency of Probit and One-hot) *Suppose G is a nearly-disconnected graph and let L be a graph Laplacian on G. Define the matrix $C = \tau^{2\alpha}(L + \tau^2 I)^{-\alpha}$ with parameters $\tau^2, \alpha > 0$. Assume $S(u^\dagger)$ is constant on the components $\tilde{G}_k$ and at least one label is observed in each component $\tilde{G}_k$. Then the probit and one-hot formulations are asymptotically consistent for any sequence $(\epsilon, \tau, \mathrm{std}(\eta)) \downarrow 0$ along which $\epsilon = o(\tau^2)$.*

Remark 4 *Conceptually the parameter ϵ should be thought of as an inherent measure of how clustered the unlabelled data is; in this paper we consider a specific set-up in which ϵ is defined as a measure of the size of edge weights between clusters. We also connect the labelled and unlabelled data via a model involving an unobserved latent variable, perturbed by noise η. Our consistency results are proven in the setting in which ϵ and $\mathrm{std}(\eta))$ both tend to zero. This is a strong assumption which, whilst allowing a precise theory, may be difficult to apply directly in practice. We believe that similar consistency results will hold under different modeling assumptions which characterize clustering and label noise in more general ways. Furthermore our consistency results demonstrate the importance of choosing the hyperparameter τ in a data-dependent fashion. Small ϵ induces a spectral gap in L and for this to translate into a spectral gap in C we require τ to be small too. However we also require $\epsilon = o(\tau^2)$ so that the number of eigenvalues of C which are at, or close to, 1 is the same as the number of clusters in the data. We also give theory and numerical evidence showing that when $\epsilon = \Theta(\tau^2)$ consistency may be lost. Our theoretical results are asymptotic in nature and therefore cannot apply directly to any one given data set. However our analysis provides insights into both algorithmic parameters choices, and algorithmic performance, in practical non-asymptotic set-ups. Indeed the papers of Bertozzi and Flenner (2012); Bertozzi et al. (2018) demonstrate the use of optimization methodologies of the type introduced here in practical non-asymptotic set-ups for real data problems, while Chen et al. (2018); Qiao et al. (2019) demonstrate analogous set-ups for related Bayesian approaches. An important conclusion of the theory and numerical experiments is that careful choice of the parameter τ is crucial for effective SSL. The take-home message here is that the use of hierarchical*

Bayesian methods, which tune τ automatically to the data, can be beneficial; as demonstrated in practical experiments in Chen et al. (2018). $\diamond$

Formal statement and proof of the preceding main theorem is given in Theorem 20 (together with Corollaries 19 and 21) for the probit method and in Theorem 32 (together with Corollary 33) for the one-hot method.

As a secondary result, accompanying the preceding theorem, we identify a natural dimension reduction for probit and one-hot optimization problems. More precisely, we show that finding $u^* \in H = \{u : Z \mapsto \mathbb{R}^M\}$ is equivalent to a similar optimization problem for a function $b^* \in H' = \{b : Z' \mapsto \mathbb{R}^M\}$. Thus we can reduce the size of the optimization problems from $N \times M$ to $J \times M$. This result, which is a discrete representer theorem, has significant practical consequences when $J \ll N$.

Theorem 5 (Dimension Reduction for Probit and One-hot) *The problem of finding u^* is equivalent to an optimization problem of the form*

$$b^* := \arg\min_{b \in H'} \frac{1}{2}\langle b, (C')^{-1} b\rangle_{H'} + \Phi'(b; y),$$

where C' is a submatrix of C after restriction of rows and columns to Z', and Φ' is defined from Φ.

A formal statement and proof of this theorem is presented in Corollary 11 for probit and Proposition 27 for the one-hot method. These results also provide identities that relate b^* to u^* and vice versa. More precisely, pointwise values of the functions b^* and u^* coincide on the labelled set Z'. Conversely, u^* can be viewed as a smooth extension of b^* from the labelled set Z' to the entire index set Z.

Finally we perform numerical experiments to illustrate the behavior of probit and one-hot methods beyond the theoretical setting. In particular we demonstrate that when $\epsilon = \Theta(\tau^2)$ these methods are not always consistent. An interesting observation we make is a sharp phase transition in the accuracy of both methods. More precisely, we observe a curve in the $(\epsilon/\tau^2, \alpha)$-plane across which the probit and one-hot methods transition rapidly from being consistent into inconsistent solutions based on majority label propagation, i.e., labelling all points in the data set according to the label that is observed most often (see Figures 2 and 8). Intuitively this happens because, for larger values of ϵ/τ^2, it is cheaper to minimize the quadratic regularization term in the optimization problem of Theorem 5 than to minimize the misfit term Φ.

1.4. Outline

Section 2 is devoted to analysis of the probit method where $M = 2$. The problem is formulated as inference for a latent real-valued function on the nodes of a graph, with the sign determining the assignation of a binary label. An optimization approach is employed in which the graph Laplacian constructed from the unlabelled data is used for regularization, and a generic zero-mean log-concave label measurement noise is assumed; this results in a convex data misfit term. We study the properties of this optimization problem, showing that the related optimization functional is convex. We prove a representer theorem and then

study asymptotic consistency of the method in Corollary 19, Theorem 20 and Corollary 21, the precise statements of Theorem 3 in the probit case.

Section 3 has the same structure as Section 2 but focuses on the multi-class setting (i.e., $M \geq 2$) and employs the one-hot method to link a real-vector-valued latent variable to the labels. The key results here are Corollary 31, Theorem 32 and Corollary 33, the precise versions of Theorem 3 in case of the one-hot method. Section 4 contains numerical experiments confirming the key theoretical results from the two preceding sections, and illustrates the behavior of probit and one-hot methods beyond the theoretical setting. In Section 5 we summarize and discuss future work.

1.5. Notation

Throughout we use Z to denote the nodes of a graph carrying a pre-assigned unlabelled data point at each node, and Z' the subset of nodes which also carry a label. We use $G_0 = \{X, W_0\}$ to denote a disconnected graph with K disconnected subgraphs (clusters) $\tilde{G}_k = \{\tilde{X}_k, \tilde{W}_k\}$ for $k = 1, \dots, K$. The $\tilde{X}_k$ are a subset of the points in X with indices $\tilde{Z}_k \subset Z$ while $\tilde{W}_k$ are submatrices of W_0. We also use $\tilde{Z}'_k$ to denote the subset of labelled points within $\tilde{Z}_k$. Subsequently we denote the graph Laplacian matrices of the subgraphs $\tilde{G}_k$ by $\tilde{L}_k$. We also introduce a nearly-disconnected graph $G_\epsilon = \{X, W_\epsilon\}$ with the weight matrix W_ϵ that is considered to be a perturbation of W_0 and use L_ϵ to denote the graph Laplacian on this nearly-disconnected graph. These concepts are introduced and discussed in subsection 2.5 and used extensively in the rest of the article.

We use u to denote real-(vector)-valued functions on Z which are acted upon by a non-linear classifier to assign labels. We use $|\cdot|$ to denote the cardinality of a set; $\langle \cdot, \cdot \rangle, \|\cdot\|$ denote the Euclidean inner-product and norm unless stated otherwise. We employ the standard Θ, $\mathcal{O}$ and o notations as in Cormen et al. (2009): given positive functions $f(s), g(s)$, we write

- $f(s) = \Theta(g(s))$ if there exist constants $c_1, c_2, s_0 > 0$ so that

$$0 \leq c_1 g(s) \leq f(s) \leq c_2 g(s) \qquad \forall\, s \in (0, s_0]\,,$$

- $f(s) = \mathcal{O}(g(s))$ if there exists $c, s_0 > 0$ so that

$$0 \leq f(s) \leq c g(s) \qquad \forall\, s \in (0, s_0]\,,$$

- $f(s) = o(g(s))$ if for any constant $c > 0$ there exists $s_0(c) > 0$ so that

$$0 \leq f(s) < c g(s) \qquad \forall\, s \in (0, s_0(c)]\,.$$

2. Binary Classification: The Probit Method

In Subsection 2.1 we set up the probit methodology, noting that the binary classification problem ($M = 2$) can be formulated using a latent variable function which is $\mathbb{R}^{M-1}$–valued rather than $\mathbb{R}^M$–valued. In Subsection 2.2 we study the likelihood contribution to the optimization problem, resulting from the labelled data, and in Subsection 2.3 the quadratic regularization resulting from the unlabelled data. In Subsection 2.4 we study the probit minimization problem, formulating the results via a discrete representer theorem, and in

Subsection 2.5 we study the properties of the representers via the properties of the eigenstructure of the covariance, exploiting the nearly-disconnected graph structure. Subsection 2.6 concludes the analysis of the probit method, studying consistency in detail.

2.1. Set-Up

We start with the case of binary classification where the nodes Z belong to only two classes. For simplicity we assume that $l(j) \in \{-1, +1\}$ for all $j \in Z$ rather than taking $l(j) \in \{1, 2\}$. This assumption is at odds with our notation in Subsection 1.2 but allows for a simpler formulation of Problem 3. Since the classes are identified with the integers $+1$ and -1 a natural choice for the classifier function S is the sign function:

$$S : \mathbb{R} \mapsto \{-1, +1\}, \qquad S(t) = \operatorname{sgn}(t) := \begin{cases} +1, & \text{if } t \geq 0, \\ -1, & \text{if } t < 0. \end{cases} \tag{5}$$

With the above choice for S we can take the latent variable u to be a real valued function on Z, i.e., $u : Z \mapsto \mathbb{R}$. We can then naturally identify the function u with a vector $\mathbf{u} \in \mathbb{R}^N$ where $\mathbf{u} = (u_1, u_2, \cdots, u_N)^T$ and $u_j = u(j)$ for $j \in Z$. This allows to view Problems 2 and 3 as the inverse problem of finding a vector $\mathbf{u}^*$ in $\mathbb{R}^N$. In the remainder of this section we will utilize this vector notation for convenience.

2.2. The Probit Likelihood

Let us begin by deriving the likelihood potential $\Phi(\mathbf{u}; y)$ for the probit method. Let S be as in (5) and recall (1), then

$$y(j) = \operatorname{sgn}(u_j + \eta_j), \qquad \eta_j \overset{iid}{\sim} \psi, \quad j \in Z,$$

wherein we have identified the noise η with a vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)^T \in \mathbb{R}^N$. Suppose ψ is a symmetric probability density function on $\mathbb{R}$ and denote the the cumulative distribution function (CDF) of ψ by Ψ. Then,

$$\mathbb{P}(y(j) = +1 | u_j) = \mathbb{P}(-u_j \leq \eta_j) = \mathbb{P}(-u_j y(j) \leq \eta_j) = \Psi(u_j y(j)).$$

For more details on this calculation, see similar arguments for the multi-class case in section 3.2. Similarly,

$$\mathbb{P}(y(j) = -1 | u_j) = \mathbb{P}(-u_j > \eta_j) = \mathbb{P}(u_j y(j) > \eta_j) = \Psi(u_j y(j)).$$

From (4) it follows that the *probit likelihood* potential $\Phi(u; y)$ has the form

$$\Phi(\mathbf{u}; y) = -\sum_{j \in Z'} \log \Psi(u_j y(j)). \tag{6}$$

2.3. Quadratic Regularization via Graph Laplacians (Binary Case)

Let us now formulate a quadratic regularization term for the probit method. Recall our encoding of the nodes Z and their similarities via a weighted graph $G = \{X, W\}$ with vertices

at x_j and edge weights $w_{ij} = w_{ji}$ for $i, j \in Z$. We denote by d_i the degree of each node $i \in Z$ as

$$d_i := \sum_{j \in Z} w_{ij},$$

and further define the diagonal matrix $D := \mathrm{diag}(d_i) \in \mathbb{R}^{N \times N}$. Finally, given constants $p, q \in \mathbb{R}$ we define the graph Laplacian operator on G

$$L := D^{-p}(D - W)D^{-q} \in \mathbb{R}^{N \times N}. \tag{7}$$

Different choices of p and q result in different normalizations of the graph Laplacian, see Belkin and Niyogi (2007); Chung (1997); Coifman and Lafon (2006); Dunlop et al. (2020); García-Trillos et al. (2019); García-Trillos and Slepčev (2018); Shi and Malik (2000); Slepčev and Thorpe (2019); Von Luxburg (2007); Von Luxburg et al. (2008) and the references therein as well as (Hoffmann et al., 2020a, Sec. 5) where a detailed discussion around various weightings of graph Laplacians and their connection to a family of elliptic operators is laid out. For example, $p = q = 0$ leads to the usual *unnormalized* graph Laplacian, when $p = q = 1/2$ we obtain the *symmetric normalized* graph Laplacian, and $p = 1$ and $q = 0$ gives the *random walk* graph Laplacian. Different normalizations of the graph Laplacian have been used for spectral clustering in the literature, but a thorough understanding of the advantages and disadvantages of certain parameter choices is still lacking, see Von Luxburg (2007). Throughout we enforce $p = q$ in order to make L symmetric with respect to the Euclidean inner-product, making no other assumptions regarding the value of p, q; however our results can be generalized to $p \neq q$ by using appropriate $D-$weighted inner-products. For $p = q$, we can then write for any vector $\mathbf{x} \in \mathbb{R}^N$,

$$\langle \mathbf{x}, L\mathbf{x} \rangle = \frac{1}{2} \sum_{i,j=1}^{N} w_{ij} \left| \frac{\mathbf{x}_i}{d_i^p} - \frac{\mathbf{x}_j}{d_j^p} \right|^2 . \tag{8}$$

Given a graph Laplacian L and parameters $\alpha, \tau^2 > 0$ we define a family of covariance operators

$$C_\tau = \tau^{2\alpha}(L + \tau^2 I)^{-\alpha} \in \mathbb{R}^{N \times N}, \tag{9}$$

where $I \in \mathbb{R}^{N \times N}$ denotes the identity matrix. We then use this covariance matrix to define the quadratic regularization term in Problem 3. To this end note that in the binary case we may identify $H = \mathbb{R}^N$; we make this identification in what follows in this section, and $\langle \cdot, \cdot \rangle$ then denotes the standard Euclidean inner-product.

Remark 6 *We use the term covariance operator to refer to the matrix C_τ following the connection between optimization problems of the form* (3) *and MAP estimators within the Bayesian formulation of probit given in Bertozzi et al. (2018). In the Bayesian perspective C_τ is the covariance operator of a Gaussian prior measure on $\mathbf{u}$, and $\mathbf{u}^*$ coincides with the MAP estimator of $\mathbf{u}^\dagger$. We note that the covariance C_τ may be viewed as a form of discrete Matérn covariance, in the framework of Lindgren et al. (2011). The scaling of C_τ that we adopt ensures that the spectrum of C_τ lies in $[0, 1]$ and hence controls the total variance of samples u from the prior: $\mathbb{E}^{N(0,C_\tau)}\|u\|^2 \approx K$, where K is the number of clusters in the disconnected or nearly-disconnected graph setting. Further study of the connections*

between C_τ and Matérn kernels is outside the scope of this article and is postponed to our companion papers (Hoffmann et al., 2020a,b) where the continuum limits of graph Laplacian and covariance matrices such as C_τ are studied when the elements of X are drawn i.i.d. at random from a probability measure, and $N \to \infty$. ◇

2.4. Properties of The Probit Minimizer

With the likelihood Φ and covariance matrix C_τ identified we can now discuss properties of the *probit functional*

$$\mathsf{J}(\mathbf{u}) := \frac{1}{2}\langle \mathbf{u}, C_\tau^{-1}\mathbf{u}\rangle + \Phi(\mathbf{u}; y), \qquad \mathbf{u} \in \mathbb{R}^N. \tag{10}$$

Remark 7 *In the following we will study the problem of minimizing J. We highlight the fact that related optimization problems for objective functions of the form*

$$\mathsf{J}(\mathbf{u}) := \frac{1}{2}\langle \mathbf{u}, L\mathbf{u}\rangle + \Upsilon(\mathbf{u}; y), \qquad \mathbf{u} \in E \tag{11}$$

have been defined and studied in Bertozzi and Flenner (2012); Bertozzi et al. (2018); Zhu et al. (2003a), although asymptotic consistency has not been investigated there. In order to give these methods a Bayesian interpretation we need to define a covariance $C = L^{-1}$, noting that L is invertible on the set $E = \{\mathbf{u} \in \mathbb{R}^N : \langle \mathbf{u}, D_0^p \mathbf{1}\rangle = 0\}$, where $\mathbf{1} \in \mathbb{R}^N$ denotes the vector of ones. L is invertible on E because $D_0^p\mathbf{1}$ spans the null-space of L when the graph is pathwise connected. Introduction of C_τ with $\tau > 0$ not only circumvents the need to work on E but also allows for consistent prior modeling of the situation in which multiple clusters have the same prior label. Furthermore the parameter α is needed when the large data limit $N \to \infty$ is considered; see Hoffmann et al. (2020a). ◇

Our first task is to prove existence and uniqueness of the minimizers of J by proving it is strictly convex. The following proposition follows directly from Bagnoli and Bergstrom (2005, Thm. 1) and states that the CDF of a log-concave probability distribution function (PDF) is also log-concave.

Proposition 8 (Convexity of the Likelihood Potential Φ) *Let ψ be a continuously differentiable, symmetric and strictly log-concave PDF with full support on $\mathbb{R}$. Then Ψ is also strictly log-concave and so $\Phi(\cdot; y) : \mathbb{R}^N \mapsto \mathbb{R}$ is strictly convex.*

Convexity of the quadratic regularization term in (10) follows directly from Lemma 34 that establishes that the matrix C_τ is strictly positive-definite whenever $\tau^2, \alpha > 0$. With the convexity of both terms in the definition of J established we can now characterize its minimizer.

Proposition 9 (Representer Theorem for the Probit Functional) *Let $G = \{X, W\}$ be a weighted graph and let ψ be a PDF that is continuously differentiable, symmetric and strictly log-concave with full support on $\mathbb{R}$. Suppose the likelihood potential Φ is given by* (6) *and the matrix C_τ is given by* (9) *with parameters $\tau^2, \alpha > 0$. Then the following hold.*

(i) The probit functional J *has a unique minimizer* $\mathbf{u}^* \in \mathbb{R}^N$.

(ii) The minimizer $\mathbf{u}^*$ *satisfies the Euler-Lagrange (EL) equations*

$$C_\tau^{-1}\mathbf{u}^* = \sum_{j\in Z'} F_j(u_j^*)\mathbf{e}_j, \tag{12}$$

where $\mathbf{e}_j$ *is the* j*-th standard coordinate vector in* $\mathbb{R}^N$ *and*

$$F_j(s) := \frac{y(j)\psi(sy(j))}{\Psi(sy(j))}. \tag{13}$$

(iii) The minimizer $\mathbf{u}^*$ *has a sparse representation*

$$\mathbf{u}^* = \sum_{j\in Z'} \breve{a}_j \mathbf{c}_j, \tag{14}$$

where $C_\tau \mathbf{e}_j =: \mathbf{c}_j = (c_{1j}, \cdots, c_{Nj})^T$ *are a subset of the column space of* $C_\tau = (c_{ij})_{i,j\in Z}$ *and* $\breve{a}_j \in \mathbb{R}$.

(iv) The vector $\mathbf{u}^*$ *defined in* (14) *solves* (12) *if and only if the coefficients* $\breve{a}_j$ *satisfy the non-linear system of equations*

$$\breve{a}_j = F_j\left(\sum_{k\in Z'} \breve{a}_k c_{jk}\right), \qquad \forall j \in Z'.$$

Proof (i) Since Ψ is the CDF of a random variable on $\mathbb{R}$ with full support then $\Psi(s) \in (0,1)$. Thus, $-\log\Psi \geq 0$ and so Φ is bounded from below. Furthermore, Φ is convex following Proposition 8. On the other hand, the matrix C_τ^{-1} is strictly positive definite following Lemma 34 and so the quadratic term $\frac{1}{2}\langle \mathbf{w}, C_\tau^{-1}\mathbf{w}\rangle$ is strictly convex and positive. Thus, since the functional J is bounded from below and is the sum of strictly convex functions then J is strictly convex and has a unique minimizer.

(ii) Since ψ is $C^1(\mathbb{R})$ the CDF Ψ is $C^2(\mathbb{R})$ and ψ/Ψ is $C^1(\mathbb{R})$ and locally bounded since ψ has full support. Then $\mathsf{J} : \mathbb{R}^N \mapsto \mathbb{R}$ is differentiable and the minimizer $\mathbf{u}^*$ satisfies the first order optimality condition $\nabla \mathsf{J}(\mathbf{u}^*) = 0$. The statement now follows by directly computing the gradient of $\mathsf{J}(\mathbf{u})$ with respect to $\mathbf{u}$.

(iii–iv) Multiply (12) by C_τ to get

$$\mathbf{u}^* = \sum_{j\in Z'} F_j(u_j^*) C_\tau \mathbf{e}_j = \sum_{j\in Z'} \breve{a}_j \mathbf{c}_j,$$

where we set $\breve{a}_j = F_j(u_j^*)$ for $j \in Z'$. Now substitute the expansion of $\mathbf{u}^*$ into the definition of $\breve{a}_j$ to get

$$\breve{a}_j = F_j\left(\left(\sum_{k\in Z'} \breve{a}_k \mathbf{c}_k\right)_j\right), \tag{15}$$

This establishes the "only if" statement in (iv). In order to establish the converse, suppose the $\breve{a}_j$ satisfy (15). Multiply this equation by $\mathbf{c}_k$ and sum over $j \in Z'$ to get

$$\sum_{j\in Z'} \breve{a}_j \mathbf{c}_j = \sum_{j\in Z'} F_j\left(\sum_{k\in Z'} \breve{a}_j c_{jk}\right)\mathbf{c}_j.$$

now define $\mathbf{u}^* = \sum_{j\in Z'} \breve{a}_j \mathbf{c}_j$ to get

$$\mathbf{u}^* = \sum_{j\in Z'} F_j(u_j^*)\mathbf{c}_j.$$

The claim follows by multiplying this equation by C_τ^{-1}. ■

Remark 10 (Connection to Kernel Regression) *We note that Proposition 9 is closely related to the representer theorem in Gaussian process and kernel regression see Rasmussen and Williams (2006, Sec. 6.2). Similar result to ours can also be found in Smola and Schölkopf (1998, Thm. 1) and Schölkopf and Smola (2002).* ◇

Part (iv) of Proposition 9 suggests that the problem of minimizing J is analogous to a low-dimensional optimization problem. To this end we now define a one-to-one reordering

$$\pi : Z' \mapsto \{1, 2, \cdots, J\}, \qquad \pi^{-1} : \{1, 2, \cdots, J\} \mapsto Z', \tag{16}$$

that allows us to associate the coefficients $\{\breve{a}_j\}_{j\in Z'}$ with a vector $\mathbf{a} = (a_1, \cdots a_J)^T \in \mathbb{R}^J$ via

$$a_{\pi(j)} = \breve{a}_j, \qquad j \in Z',$$

and define submatrix $C' \in \mathbb{R}^{J\times J}$ by the identity

$$(C'_\tau)_{\pi(i),\pi(j)} = c'_{ij}, \qquad i, j \in Z'. \tag{17}$$

That is, C'_τ is the matrix C_τ with the rows and columns of the indices in $Z \setminus Z'$ removed. Finally, we define $\mathbf{b} := C'_\tau \mathbf{a}$. We then have the following natural dimension reduction for the probit optimization problem.

Corollary 11 (Probit Dimension Reduction) *Suppose the conditions of Proposition 9 are satisfied. Then the following hold.*

(i) The problem of finding the minimizer $\mathbf{u}^* \in \mathbb{R}^N$ of the functional J is equivalent to the problem of finding the vector $\mathbf{b}^* \in \mathbb{R}^J$ that solves

$$(C'_\tau)^{-1}\mathbf{b}^* = F'(\mathbf{b}^*), \tag{18}$$

where the map $F' : \mathbb{R}^J \mapsto \mathbb{R}^J$ is defined as

$$F'(\mathbf{v}) = (f_1(v_1), \cdots, f_J(v_J))^T, \qquad f_k(v_k) := F_{\pi^{-1}(k)}(v_k)$$

and F_j are defined in (13).

(ii) Moreover, the vector $\mathbf{b}^*$ solves the optimization problem

$$\mathbf{b}^* = \operatorname*{arg\,min}_{\mathbf{v}\in\mathbb{R}^J} \mathsf{J}'(\mathbf{v}),$$

where

$$\mathsf{J}'(\mathbf{b}) := \frac{1}{2}\langle \mathbf{b}, (C'_\tau)^{-1}\mathbf{b}\rangle + \Phi'(\mathbf{b}; y),$$

and

$$\Phi'(\mathbf{b}; y) = -\sum_{j=1}^{J} \log \Psi\Big(b_j y(\pi^{-1}(j))\Big).$$

(iii) The two solutions $\mathbf{b}^* \in \mathbb{R}^J$ *and* $\mathbf{u}^* \in \mathbb{R}^N$ *satisfy the relationship*

$$\mathbf{u}^* = \sum_{j \in Z'} \left((C'_\tau)^{-1} \mathbf{b}^* \right)_{\pi(j)} \mathbf{c}_j, \tag{19}$$

and

$$b_k^* = u^*_{\pi^{-1}(k)}, \qquad k = \{1, 2, \cdots, J\}.$$

Proof This result follows from Proposition 9 and direct computations. ■

Remark 12 (Variable Elimination and Gaussian Process Regression) *There is a simple explanation for the finite-dimensional representer theorem which underlies Proposition 9 and Proposition 11. If we re-order the variables in* $\mathbf{u}$ *into components* $\mathbf{u}^+$ *in* Z' *and* $\mathbf{u}^-$ *in* $Z \setminus Z'$*, and re-order the components of the precision matrix* $P = C_\tau^{-1}$ *then setting the gradient of* J *to* 0 *in this re-ordered set of variables gives equations of the form*

$$\begin{pmatrix} P^{++} & P^{+-} \\ P^{-+} & P^{--} \end{pmatrix} \begin{pmatrix} \mathbf{u}^+ \\ \mathbf{u}^- \end{pmatrix} = \begin{pmatrix} \mathbf{g}(\mathbf{u}^+) \\ 0 \end{pmatrix}.$$

This follows from the fact that $\Phi(\mathbf{u})$ *does not depend on* $\mathbf{u}^-$*; the term* $\mathbf{g}(\mathbf{u}^+)$ *results from the gradient of* $\Phi(\mathbf{u})$ *with respect to* $\mathbf{u}^+$*. From this re-ordering of the equations several things are apparent: (i) the bottom row provides a linear mapping from* $\mathbf{u}^+$ *to* $\mathbf{u}^-$ *since* P^{--} *is invertible whenever* C_τ *is; (ii) using this linear mapping it is possible to obtain a closed nonlinear equation for* $\mathbf{u}^+$ *only, from the top row, and the linear part of this equation has a Schur complement form; (iii) the unknown* $\mathbf{u}^-$ *is recovered by solving a linear equation; (iv) the nonlinear equation for* $\mathbf{u}^+$ *may be viewed as the equation for a critical point of a functional of* $\mathbf{u}^+$ *only. These four points are encapsulated in the previous results, where they are rendered in a form familiar from Gaussian process regression and representer theorems. Ideas analogous to those described in this remark underlie all representer theorems, but are not so transparent in the infinite-dimensional setting. We present the results in the abstract form of Proposition 9 and Corollary 11 to highlight the formal analogies with our companion papers Hoffmann et al. (2020a,b) which study the limiting continuum optimization problems that arise in the* $N \to \infty$ *limit.* ◇

The expansion (14) indicates that the minimizer $\mathbf{u}^* \in \text{span } \{\mathbf{c}_j\}_{j \in Z'}$; we refer to the $\mathbf{c}_j$ as representers. In other words, the minimizer $\mathbf{u}^*$ belongs to a subspace of the column space of the covariance matrix C_τ. Recall that by definition $C_\tau = \tau^{2\alpha}(L + \tau^2 I)^{-\alpha}$ and so we can compute the vectors $\mathbf{c}_j$ by solving the linear equations,

$$(L + \tau^2 I)^\alpha \mathbf{c}_j = \tau^{2\alpha} \mathbf{e}_j, \qquad j \in Z', \tag{20}$$

that cost J linear solves involving an $N \times N$ matrix. With the $\{\mathbf{c}_j\}_{j \in Z'}$ at hand we can extract the matrix C'_τ and solve the nonlinear system (18) for $\mathbf{b}^*$ and in turn compute the solution $\mathbf{u}^*$ by (19); note that F_j is defined in Proposition 9(ii). Then, whenever $J \ll N$ solving the dimension reduced problem (18) is typically much faster than solving the full nonlinear system (12). We present evidence of this improved efficiency in Subsection 4.3 in the context of the one-hot method for multi-class classification.

We now exploit the geometry in the problem dictated by the nearly-disconnected graph structure that forms the basic assumption underlining our consistency analysis. It is clear that the geometry of $\mathbf{u}^*$ is dictated by the geometry of the vectors $\mathbf{c}_j$. It is then natural for us to try to identify the geometry of the $\mathbf{c}_j$. By Lemma 35 we have the expansion

$$\mathbf{c}_j = \sum_{k=1}^{N} \frac{1}{\lambda_k} (\boldsymbol{\phi}_k)_j \boldsymbol{\phi}_k, \tag{21}$$

where $\{\lambda_k, \boldsymbol{\phi}_k\}$ are the eigenpairs of C_τ^{-1}. Therefore, by analyzing the spectrum of C_τ we can identify the geometry of the vectors $\mathbf{c}_j$ which together with the vector $\mathbf{b}^*$ allow us to identify the minimizer $\mathbf{u}^*$ and eventually prove consistency of the probit minimizer. Spectral analysis of C_τ is outlined in Appendix A, and in the next subsection we present the main propositions and assumptions that are used in the remainder of the article.

2.5. Perturbation Theory for Covariance Operators

Consider a disconnected graph $G_0 = \{X, W_0\}$ consisting of $K < N$ connected components $\tilde{G}_k$, i.e., the subgraphs $\tilde{G}_k$ are connected but there exist no edges between pairs of components $\tilde{G}_i$, $\tilde{G}_k$ with $i \neq k$. Without loss of generality assume the nodes in Z are ordered so that $Z = \{\tilde{Z}_1, \tilde{Z}_2, \cdots, \tilde{Z}_K\}$ and the $\tilde{Z}_k$ collect the nodes in the k-th subgraph $\tilde{G}_k$. We refer to the $\tilde{Z}_k$ as *clusters*. Thus, the weight matrix $W_0 = (w_{ij}^{(0)})$ satisfies

$$\begin{cases} w_{ij}^{(0)} \geq 0 & \text{if } i \neq j \text{ and } i, j \in \tilde{Z}_k \text{ for some } k, \\ w_{ij}^{(0)} = 0 & \text{if } i = j \text{ or } i \in \tilde{Z}_k, j \in \tilde{Z}_\ell, \text{ for } k \neq \ell. \end{cases} \tag{22}$$

We will show that when τ is small the geometry of $\mathbf{c}_j$ is dominated by indicator functions of the clusters $\tilde{Z}_k$. First, let us collect some assumptions on the graph G_0.

Assumption 1 *The graph $G_0 = \{X, W_0\}$ satisfies the following conditions with $K < N$:*

(a) The weight matrix W_0 satisfies (22) and has a block diagonal form $W_0 = diag(\tilde{W}_1, \cdots, \tilde{W}_K)$ where $\tilde{W}_k$ are the weight matrices of the subgraphs $\tilde{G}_k$.

(b) Let $\tilde{L}_k$ be the graph Laplacian matrices of the subgraphs $\tilde{G}_k$, i.e.,

$$\tilde{L}_k := \tilde{D}_k^{-p} (\tilde{D}_k - \tilde{W}_k) \tilde{D}_k^{-p}$$

with $\tilde{D}_k$ denoting the degree matrix of $\tilde{W}_k$. There exists a uniform constant $\theta > 0$ so that for $j = 1, \cdots, K$ the submatrices $\tilde{L}_j$ have a uniform spectral gap, i.e.,

$$\langle \mathbf{x}, \tilde{L}_j \mathbf{x} \rangle \geq \theta \langle \mathbf{x}, \mathbf{x} \rangle, \tag{23}$$

for all vectors $\mathbf{x} \in \mathbb{R}^{N_k}$ and $\mathbf{x} \perp \tilde{D}_k^p \mathbf{1}_k$ where $\mathbf{1}_k \in \mathbb{R}^{N_k}$ are vectors of ones.

Note that the preceding assumption means that the clusters $\tilde{G}_k$ are pathwise connected. Further, condition (23) excludes the possibility of outliers, that is, nodes of zero degree. This means that the inner product as expressed in (8) is well defined. In the following

and throughout the remainder of the article we introduce the following notation: We define the graph Laplacian in terms of the weight matrix W_0 and the associated degree matrix $D_0 := \mathrm{diag}(d_i^{(0)})$:

$$L_0 := D_0^{-p}(D_0 - W_0)D_0^{-p} \in \mathbb{R}^{N\times N}. \tag{24}$$

Let $Z = \{\tilde{Z}_1, \tilde{Z}_2, \cdots, \tilde{Z}_K\}$ and define the weighted indicator functions

$$(\boldsymbol{\chi}_k)_j := \begin{cases} \left(d_j^{(0)}\right)^p, & \text{if } j \in \tilde{Z}_k, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \bar{\boldsymbol{\chi}}_k := \frac{1}{\|\boldsymbol{\chi}_k\|}\boldsymbol{\chi}_k. \tag{25}$$

Similarly, the weighted indicator function on Z is denoted by

$$\boldsymbol{\chi} := D_0^p \mathbf{1}, \qquad \text{and} \qquad \bar{\boldsymbol{\chi}} := \frac{1}{\|\boldsymbol{\chi}\|}\boldsymbol{\chi}. \tag{26}$$

The next proposition, whose proof is given in Appendix A.1, identifies the geometry of the covariance matrix constructed from L_0. The key take away is that, for small τ, the covariance matrix is nearly block diagonal which implies negligible correlation outside clusters.

Proposition 13 *Let $G_0 = \{X, W_0\}$ satisfy Assumption 1 and let L_0 be a graph Laplacian of form (24) on G_0 and define the covariance matrix $C_{\tau,0}$ on G_0 for $\tau^2, \alpha > 0$*

$$C_{\tau,0} := \tau^{2\alpha}(L_0 + \tau^2 I)^{-\alpha}. \tag{27}$$

Then as $\tau \downarrow 0$,

$$\left\|\mathbf{c}_{j,0} - (\bar{\boldsymbol{\chi}}_k)_j \bar{\boldsymbol{\chi}}_k\right\|^2 \le \Xi\tau^{4\alpha} \qquad \forall j \in \tilde{Z}_k,$$

where $\mathbf{c}_{j,0} = (c_{1j}^{(0)}, \cdots, c_{Nj}^{(0)})^T$ is the j-th column of $C_{\tau,0}$, and $\Xi > 0$ is a uniform constant.

Thus, when $\tau^{2\alpha}$ is small the vectors $\mathbf{c}_{j,0}$ have a similar geometry to the set functions $\bar{\boldsymbol{\chi}}_k$. We now show that this result remains true when the graph G_0 is perturbed.

Consider a perturbation of the matrix W_0 by modifying some of the entries $w_{ij}^{(0)}$ and possibly making the graph connected. More precisely, let $G_\epsilon = \{X, W_\epsilon\}$ where

$$W_\epsilon = W_0 + \sum_{h=1}^{\infty} \epsilon^h W^{(h)}. \tag{28}$$

We need to collect some assumptions on the perturbed matrix W_ϵ to restrict the type of perturbations that are allowed.

Assumption 2 *The graph $G_\epsilon = \{X, W_\epsilon\}$ satisfies the following assumptions:*

(a) The weight matrix W_ϵ satisfies expansion (28) with W_0 satisfying (22).

(b) The sequence of matrix norms satisfies $\{\|W^{(h)}\|_2\}_{h\in\mathbb{Z}} \in \ell^\infty$, and for each $h \in \mathbb{Z}$, $W^{(h)} = (w_{ij}^{(h)})$ is self-adjoint and satisfies

$$\left\{ w_{ij}^{(h)} \ge 0, \quad \text{if} \quad w_{ij}^{(0)} = 0 \quad \text{for} \quad i, j \in Z, i \ne j, \right. \tag{29}$$

Note that $w_{ij}^{(h)}$ may be negative for indices i, j such that $w_{ij}^{(0)} > 0$.

Associated to the weight matrix W_ϵ is a graph Laplacian and covariance matrix

$$L_\epsilon := D_\epsilon^{-p}(D_\epsilon - W_\epsilon)D_\epsilon^{-p}, \qquad C_{\tau,\epsilon} := \tau^{2\alpha}(L_\epsilon + \tau^2 I)^{-\alpha}, \tag{30}$$

where $\tau^2, \alpha > 0$ and $p \in \mathbb{R}$. We then have the following result stating that if $\epsilon = o(\tau^2)$, then the geometry of the column space of $C_{\tau,\epsilon}$ is close to the set functions $\boldsymbol{\chi}_k$, whilst if $\epsilon = \Theta(\tau^2)$ then prior correlation between clusters is introduced; see Appendix A.2 for the proof.

Proposition 14 *Suppose G_0 satisfies Assumption 1 and G_ϵ satisfies Assumption 2. For $\tau^2, \alpha > 0$ define the covariance matrix $C_{\tau,\epsilon}$ as in (30) and denote the j-th column of $C_{\tau,\epsilon}$ by $\mathbf{c}_{j,\epsilon} = (c_{1j}^{(\epsilon)}, \cdots, c_{Nj}^{(\epsilon)})^T$. Then*

(a) If $\epsilon = o(\tau^2)$, then there exists a constant $\Xi > 0$ independent of ϵ and τ so that

$$\left\|\mathbf{c}_{j,\epsilon} - (\bar{\boldsymbol{\chi}}_k)_j \bar{\boldsymbol{\chi}}_k\right\|^2 \leq \Xi\left(\epsilon^2/\tau^4 + \tau^{4\alpha} + \epsilon^2\right), \qquad \forall j \in \tilde{Z}_k.$$

.

(b) If $\epsilon/\tau^2 = \beta > 0$ is constant, then there exist constants $\Xi, \Xi' > 0$, independent of ϵ and τ so that

$$\left\|\mathbf{c}_{j,\epsilon} - \left[(1-\breve{\beta})(\bar{\boldsymbol{\chi}})_j \bar{\boldsymbol{\chi}} + \breve{\beta}(\bar{\boldsymbol{\chi}}_k)_j \bar{\boldsymbol{\chi}}_k\right]\right\|^2 \leq \Xi\left(\epsilon^2 + \tau^{4\alpha}\right), \qquad \forall j \in \tilde{Z}_k,$$

and where $\breve{\beta} = (1 + \Xi'\beta)^{-\alpha}$.

2.6. Consistency of the Probit Method

Throughout this section we consider a graph $G_0 = \{X, W_0\}$ consisting of K components, along with perturbed graphs $G_\epsilon = \{X, W_\epsilon\}$ as introduced in the previous subsection. As before, we use $Z' \subset Z$ to denote the set of points where labels are observed and assume the usual ordering $Z = \tilde{Z}_1 \cup \tilde{Z}_2 \cup \cdots \cup \tilde{Z}_K$ where $\tilde{Z}_k$ denotes the k-th cluster in Z. Recall the probit assumption on the labelled data, namely that

$$y(j) = \operatorname{sgn}(u_j^\dagger + \eta_j), \qquad j \in Z', \tag{31}$$

where $\mathbf{u}^\dagger = (u_1^\dagger, \cdots, u_N^\dagger)^T$ is the vector isomorphic to the ground truth function $u^\dagger$. The additive noises η_j are assumed to be a rescaling of a sequence of i.i.d. samples from a reference density ψ. That is, for all $j \in Z'$,

$$\eta_j = \gamma\breve{\eta}_j, \qquad \breve{\eta}_j \overset{iid}{\sim} \psi, \tag{32}$$

where ψ is the PDF of a centered random variable with unit standard deviation, and thus $\gamma > 0$ is the standard deviation of the η_j. Thus the η_j are i.i.d and have distribution

$$\psi_\gamma(t) = \frac{1}{\gamma}\psi\left(\frac{t}{\gamma}\right). \tag{33}$$

We recall a useful result stating that log-concave random variables have exponential tails (Bogachev, 2010, Thm. 4.3.7).

Lemma 15 *Let $\psi(t)$ be a log-concave PDF on $\mathbb{R}$. Then there is $\omega_c > 0$ such that, for all $\omega \in [0, \omega_c)$, $\int_{\mathbb{R}} \exp(\omega|t|)\psi(t)dt < +\infty$.*

With this lemma we can estimate the probability of the event where the observed labels $y(j)$ have the same value as $\operatorname{sgn}(u_j^\dagger)$, i.e., the event where the data is exact.

Lemma 16 *Let $\psi(t)$ be a log-concave PDF on $\mathbb{R}$. Then there exist constants $\omega_1, \omega_2 > 0$ depending only on ψ, so that*

$$\mathbb{P}\left(y(j) = \operatorname{sgn}(u_j^\dagger),\ \textit{for all } j \in Z'\right) \geq \prod_{j \in Z'} \left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}|u_j^\dagger|\right)\right].$$

That is, when $\gamma > 0$ is small the data y is exact with high probability.

Proof By Lemma 15 there exists a sufficiently small $\omega_1 > 0$ and constant $\omega_2 > 0$ so that for $\gamma > 0$,

$$\omega_2 = \int_{\mathbb{R}} \exp\left(\frac{\omega_1}{\gamma}|t|\right)\psi_\gamma(t)dt < +\infty.$$

Let $\eta_j \sim \psi_\gamma$ then by Markov's inequality for $\theta > 0$

$$\mathbb{P}(\eta_j \geq \theta) \leq \omega_2 \exp\left(-\frac{\omega_1}{\gamma}\theta\right).$$

But $y(j) \neq \operatorname{sgn}(u^\dagger(j))$ whenever

$$\begin{aligned} \eta_j &< -|u_j^\dagger| \quad \text{if} \quad u_j^\dagger \geq 0, \\ \eta_j &\geq |u_j^\dagger| \quad \text{if} \quad u_j^\dagger < 0. \end{aligned}$$

The result now follows from the symmetry of the ψ_γ and independence of the η_j. ■

Lemma 17 *Suppose (31) and (32) hold, ψ is log-concave and $|u_j^\dagger| > \theta > 0$ for all $j \in Z'$. Then for any sequence $\gamma \downarrow 0$, $y(j) \xrightarrow{\text{a.s.}} \operatorname{sgn}(u_j^\dagger)$ a.s. with respect to $\prod_{j \in Z'} \psi(t_j)$ the law of the i.i.d. sequence $\{\breve{\eta}_j\}_{j \in z'}$.*

Proof Since ψ is log-concave it has exponential tails by Lemma 15 and $|\breve{\eta}_j| < \infty$ a.s. [1] Recall that value of $\eta_j = \gamma\breve{\eta}_j$. Then for any fixed $\breve{\eta}_j \in (-\infty, \infty)$, if $\gamma < \theta/|\breve{\eta}_j|$ then $y(j) = \operatorname{sgn}(u_j^\dagger)$. Since $\breve{\eta}_j$ are a.s. finite the result follows. ■

Now consider a probit likelihood potential of the form

$$\Phi_\gamma(\mathbf{u}; y) := \sum_{j \in Z'} -\log \Psi_\gamma(u_j y(j)), \tag{34}$$

where

$$\Psi_\gamma(s) = \int_{-\infty}^{s} \psi_\gamma(t)dt, \qquad s \in \mathbb{R}. \tag{35}$$

For $\epsilon, \tau^2, \gamma > 0$ we study the consistency of minimizers of the functionals

$$\mathsf{J}_{\tau,\epsilon,\gamma}(\mathbf{u}) := \frac{1}{2}\langle \mathbf{u}, C_{\tau,\epsilon}^{-1}\mathbf{u}\rangle + \Phi_\gamma(\mathbf{u}; y), \qquad \mathbf{u} \in \mathbb{R}^N. \tag{36}$$

This functional is of the same form as (10), and so the results in Section 2.4 apply.

1. In fact the proof reveals that all we need is that the $\breve{\eta}_j$ are a.s. finite, for which log-concavity suffices.

2.6.1. Probit Consistency With A Single Observed Label

We start with the simple case of a single observed label. Without loss of generality assume $Z' = \{1\}$, that is the observed label is the first point in the first cluster $\tilde{Z}_1$ and $u_1^\dagger > 0$.

Proposition 18 *Consider the single observation setting above and suppose Assumptions 1 and 2 are satisfied by G_0 and G_ϵ. Let $\mathbf{u}^*$ denote the minimizer of $\mathsf{J}_{\tau,\epsilon,\gamma}$ and let $\gamma > 0$.*

(a) If $\epsilon = o(\tau^2)$ as $\tau \to 0$ then $\exists \tau_0 > 0$ so that $\forall (\tau, \gamma) \in (0, \tau_0) \times (0, \infty)$ and $\forall j \in \tilde{Z}_1$

$$\mathbb{P}\Big(\operatorname{sgn}\left(u_j^*\right) = +1\Big) \geq 1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}|u_1^\dagger|\right), \tag{37}$$

where $\omega_1, \omega_2 > 0$ are uniform constants depending only on ψ.

(b) If $\epsilon = \Theta(\tau^2)$ then the above statement holds for all $j \in Z$.

Proof (a) By Corollary 11 we have that $u_1^* = b$ where b solves

$$b = (C_{\tau,\epsilon})_{11} F_{1,\gamma}(b). \tag{38}$$

where we recall $F_{1,\gamma}(s) = y(1)\psi_\gamma(sy(1))/\Psi_\gamma(sy(1))$. Furthermore, by Proposition 14(a) and equivalence of ℓ_∞ and ℓ_2 norms we infer that there exists $\tau_0 > 0$ so that $\forall \tau \in (0, \tau_0)$ we have

$$(\mathbf{c}_{j,\epsilon})_1 = \begin{cases} (\bar{\boldsymbol{\chi}}_1)_j\,(\bar{\boldsymbol{\chi}}_1)_1 + \mathcal{O}\left(\epsilon/\tau^2 + \tau^{2\alpha} + \epsilon\right), & j \in \tilde{Z}_1, \\ \mathcal{O}\left(\epsilon/\tau^2 + \tau^{2\alpha} + \epsilon\right), & j \notin \tilde{Z}_1. \end{cases} \tag{39}$$

Thus, we can rewrite (38) up to leading order in the form

$$b = |(\bar{\boldsymbol{\chi}}_1)_1|^2 F_{1,\gamma}(b).$$

Now consider the event where $y(1) = +1$, i.e., the measurement is exact. Then $b > 0$ since $F_{1,\gamma} > 0$ and $(\bar{\boldsymbol{\chi}}_1)_1 > 0$. Finally, by Corollary 11(iii) we can write

$$\mathbf{u}^* = F_{1,\gamma}(b)\mathbf{c}_{1,\epsilon} = F_{1,\gamma}(b)(\bar{\boldsymbol{\chi}}_1)_1\bar{\boldsymbol{\chi}}_1 + \mathcal{O}\left(\epsilon/\tau^2 + \tau^{2\alpha} + \epsilon\right).$$

Thus, when $\epsilon/\tau^2, \tau$ and ϵ are sufficiently small and the data is exact, $\mathbf{u}^*$ is positive on $\tilde{Z}_1$. The claim now follows by Lemma 16. The statement in (b) follows by an identical argument except that in this case

$$\mathbf{c}_{1,\epsilon} = \left[(1 - \breve{\beta})(\bar{\boldsymbol{\chi}})_1\bar{\boldsymbol{\chi}} + \breve{\beta}(\bar{\boldsymbol{\chi}}_1)_1\bar{\boldsymbol{\chi}}_1\right] + \mathcal{O}\left(\tau^{2\alpha} + \epsilon\right)$$

with $\breve{\beta} \in (0, 1)$. Thus in the event that $y(1) = +1$ the minimizer $\mathbf{u}^*$ is positive on all of Z. ■

The following corollary shows that the relationship between τ (which is user-specified) and ϵ (which is a property of the unlabelled data) is crucial in determining how the probit algorithm assigns labels to the entire data set in the small noise limit; case (a) is neutral about $Z \setminus \tilde{Z}_1$ whilst case (b) leads to $Z \setminus \tilde{Z}_1$ being labelled the same as $\tilde{Z}_1$. The reason for the difference is that the limit process in (a) corresponds, asymptotically, to a setting

where a priori different clusters have no correlation whilst under the limit process (b) there is positive correlation; thus, under (b), the one given label is propagated to the entire set of nodes. When more clusters are labelled then similar effects are present under the limit process (b), but are harder to express analytically because they set-up a competition between potentially conflicting prior information and observed label information. This is investigated numerically in Section 4.

Corollary 19 *Consider the single observation setting as above. Suppose Proposition 18 is satisfied and* $|u_j^\dagger| > 0$ *for all* $j \in Z'$*. Then the following holds with a.s. convergence in the sense of Lemma 17:*

(a) For any sequence $\gamma, \tau, \epsilon \downarrow 0$ *along which* $\epsilon = o(\tau^2)$ *it holds that*

$$\operatorname{sgn}\left(u_j^*\right) \xrightarrow{\text{a.s.}} \operatorname{sgn}(u_1^\dagger), \qquad \forall j \in \tilde{Z}_1.$$

(b) For any sequence $\gamma, \tau, \epsilon \downarrow 0$ *along which* $\epsilon = \Theta(\tau^2)$ *the above statement holds for all* $j \in Z$*.*

Proof In the the proof of Proposition 18(a) we showed that $\operatorname{sgn}(u_j^*) = +1$ on $\tilde{Z}_1$ so long as the data $y(1) = +1$ independent of $\gamma > 0$. In light of this, (a) follows directly from Lemma 17 implying that the data $y(j) \to \operatorname{sgn}(u_j^\dagger)$ a.s. as $\gamma \downarrow 0$. Statement (b) follows in the same way but using the proof of Proposition 18(b). ■

2.6.2. Probit Consistency with Multiple Observed Labels

Let us now consider the setting where multiple labels are observed, i.e. $|Z'| = J \geq 2$. We need to make an additional assumption on the ground truth function $\mathbf{u}^\dagger$.

Assumption 3 *Let* $\tilde{Z}_k$ *be a cluster within which a label has been observed, i.e.,* $\tilde{Z}_k \cap Z' \neq \varnothing$*. Then* $\operatorname{sgn}(\mathbf{u}^\dagger)$ *does not change within* $\tilde{Z}_k$*.*

It is helpful in the following to define Z'' to be the index set of nodes within all clusters $\tilde{Z}_k$ where a label has been observed, i.e.,

$$Z'' := \cup_{\{k:\tilde{Z}_k \cap Z' \neq \varnothing\}} \tilde{Z}_k. \tag{40}$$

Theorem 20 *Consider the multiple observation setting above and suppose Assumptions 1, 2 and 3 are satisfied by* G_0*,* G_ϵ *and* $\mathbf{u}^\dagger$*. Let* $\mathbf{u}^*$ *be the minimizer of* $\mathsf{J}_{\tau,\epsilon,\gamma}$*. If* $\epsilon = o(\tau^2)$ *as* $\tau \to 0$ *then* $\exists \tau_0 > 0$ *so that* $\forall (\tau, \gamma) \in (0, \tau_0) \times (0, \infty)$ *and* $\forall j \in Z''$

$$\mathbb{P}\Big(\operatorname{sgn}\left(u_j^*\right) = \operatorname{sgn}(u_j^\dagger)\Big) \geq \prod_{i \in Z'} \left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}|u_i^\dagger|\right)\right], \qquad \forall j \in Z'',$$

where $\omega_1, \omega_2 > 0$ *are uniform constants depending only on* ψ*.*

Proof Our proof follows a similar approach to the single observation case. The main difference is that now the dimension reduced system (18) takes the form

$$(C'_{\tau,\epsilon})^{-1}\mathbf{b}^* = F'_\gamma(\mathbf{b}^*). \tag{41}$$

Where $C'_{\tau,\epsilon}$ is now the submatrix of $C_{\tau,\epsilon}$ with the rows and columns of the indices $Z \setminus Z'$ removed. It follows from Proposition 14 that for small τ the matrix $C'_{\tau,\epsilon} = (c'_{ij})$ approaches a block diagonal matrix and so

$$c'_{ij} = \begin{cases} (\bar{\boldsymbol{\chi}}_k)_{\pi^{-1}(j)}\,(\bar{\boldsymbol{\chi}}_k)_{\pi^{-1}(i)} + \mathcal{O}\left(\epsilon/\tau^2 + \tau^{2\alpha} + \epsilon\right), & \text{if} \quad \pi^{-1}(i), \pi^{-1}(j) \in \tilde{Z}_k, \\ \mathcal{O}\left(\epsilon/\tau^2 + \tau^{2\alpha} + \epsilon\right), & \text{if} \quad \pi^{-1}(i) \in \tilde{Z}_k, \pi^{-1}(j) \in \tilde{Z}_\ell, k \neq \ell. \end{cases} \tag{42}$$

Without loss of generality assume that observations are made in the clusters $\tilde{Z}_1, \cdots, \tilde{Z}_{K'}$. Note that $K' \leq K$ since we do not need to assume observations are made in every cluster. Let $\tilde{Z}'_1, \cdots, \tilde{Z}'_{K'}$ denote the indices of the labelled nodes in the corresponding clusters and define $J_k := |\tilde{Z}'_k|$. Then (41) approximately decouples between the clusters and up to leading order we can write

$$(\bar{\boldsymbol{\chi}}_k)_j^{-1}\, b^*_{\pi(j)} = \sum_{i \in \tilde{Z}'_k} (\bar{\boldsymbol{\chi}}_k)_i\, F_{i,\gamma}(b^*_{\pi(i)}), \qquad \text{for } j \in \tilde{Z}'_k \text{ and } k \in \{1, \cdots, K'\}.$$

Observe that the right hand side is independent of j and so, writing $b^*_\ell = u^*_{\pi^{-1}(\ell)}$ for $\ell \in \{1, 2, \cdots, J\}$ as in Corollary 11(iii), it follows that $D_0^{-p}\mathbf{u}^*$ is a constant vector on the index sets $\tilde{Z}'_k$. Thus, we can further simplify this equation to get

$$(\bar{\boldsymbol{\chi}}_k)_j^{-1}\, u^*_j = \sum_{i \in \tilde{Z}'_k} (\bar{\boldsymbol{\chi}}_k)_i\, F_{i,\gamma}(u^*_j), \qquad \text{for } j \in \tilde{Z}'_k \text{ and } k \in \{1, \cdots, K'\},$$

which we only need to solve once on every cluster. Finally, observe that $\operatorname{sgn}(\mathbf{u}^\dagger)$ does not change on $\tilde{Z}_k$ following Assumption 3 and so in the event that y is exact we have

$$u^*_j = (\bar{\boldsymbol{\chi}}_k)_j \left(\sum_{i \in \tilde{Z}'_k} (\bar{\boldsymbol{\chi}}_k)_i \right) \frac{\operatorname{sgn}(u^\dagger_j)\psi_\gamma\left(\operatorname{sgn}(u^\dagger_j)u^*_j\right)}{\Psi_\gamma\left(\operatorname{sgn}(u^\dagger_j)u^*_j\right)}, \qquad \text{for } j \in \tilde{Z}'_k \text{ and } k \in \{1, \cdots, K'\}.$$

To this end, $\operatorname{sgn}(\mathbf{u}^*)$ agrees with $\operatorname{sgn}(\mathbf{u}^\dagger)$ on the observation nodes. Once again using Corollary 11(iii) we see that

$$\mathbf{u}^* = \sum_{j \in \tilde{Z}'} \frac{\operatorname{sgn}(u^\dagger_j)\psi_\gamma\left(\operatorname{sgn}(u^\dagger_j)b^*_{\pi(j)}\right)}{\Psi_\gamma\left(\operatorname{sgn}(u^\dagger_j)b^*_{\pi(j)}\right)}\mathbf{c}_{j,\epsilon} = \sum_{k=1}^{K'} \breve{a}_k \bar{\boldsymbol{\chi}}_k,$$

where the last identity is once more up to leading order following Proposition 14 and for coefficients

$$\breve{a}_k := \sum_{j \in Z'} \frac{\operatorname{sgn}(u^\dagger_j)\psi_\gamma\left(\operatorname{sgn}(u^\dagger_j)b^*_{\pi(j)}\right)}{\Psi_\gamma\left(\operatorname{sgn}(u^\dagger_j)b^*_{\pi(j)}\right)} (\bar{\boldsymbol{\chi}}_k)_j\,,$$

such that $\operatorname{sgn}(\breve{a}_k)$ agrees with $\operatorname{sgn}(\mathbf{u}^\dagger)$ on $\tilde{Z}_k$ for $k = 1, \cdots, K'$. Finally, the claim follows by applying Lemma 16 to compute the probability of the event where the data is exact. ■

Similarly to Corollary 19 the next corollary follows from the proof of Theorem 20 and Lemma 17.

Corollary 21 *Suppose Theorem 20 is satisfied and $|u_j^\dagger| > 0$ for $j \in Z'$. Then for any sequence $\gamma, \tau, \epsilon \downarrow 0$ along which $\epsilon = o(\tau^2)$ it holds that,*

$$\operatorname{sgn}\left(u_j^*\right) \xrightarrow{\text{a.s.}} \operatorname{sgn}(u_j^\dagger) \qquad \forall j \in Z'',$$

with a.s. convergence in the sense of Lemma 17.

3. Multi-Class Classification: The One-Hot Method

In the multi-class setting $M > 2$ we can no longer use the $\operatorname{sgn}(\cdot)$ function to reduce the dimension of the latent variable to $\mathbb{R}^{M-1}$ as we did for binary $M = 2$ classification in Section 2. Instead we use one-hot encoding and work directly with latent variables taking value in $\mathbb{R}^M$. We set up the one-hot methodology in Subsection 3.1 assuming that $M \geq 2$, though it would be un-necessary to use it for $M = 2$ and Gaussian label noise when it reduces to probit. In Subsection 3.2 we study the form of the one-hot likelihood that appears in the optimization problem, resulting from the labelled data. In Subsection 3.3 we introduce a quadratic regularization term for the one-hot method that uses the covariance matrix C_τ of (9) and is analogous to the quadratic penalty used in the probit method. In Subsection 3.4 we study the one-hot minimization problem, formulating a discrete representer theorem for the one-hot method. Subsection 3.5 concludes the analysis of the one-hot method, studying consistency in some detail by putting together the results of previous subsection with the spectral theory introduced in Section 2.5.

3.1. Set-Up

We now turn our attention to the multi-class classification problem, i.e., where the label function $l : Z \mapsto \{1, \cdots, M\}$ assigns one of $M \geq 1$ classes to each point in X. In this case the sign function from Section 2 is no longer an appropriate classifying function and we need a different method. We shall utilize the *one-hot mapping*

$$S(\mathbf{v}) = \arg\max_k v_k, \qquad \mathbf{v} = (v_1, \cdots, v_M) \in \mathbb{R}^M. \tag{43}$$

In the case of two maximal elements $v_{k_1} = v_{k_2}$, we take the smallest index. As with probit, the case of a near-tie is prone to misclassification by perturbation. For the purpose of consistency analysis, we later make assumptions that ensure a tie for the maximal element cannot occur.

The latent variable $u : Z \mapsto \mathbb{R}^M$ is isomorphic to a matrix $U = (u_{mj}) \in \mathbb{R}^{M \times N}$. We use $\mathbf{u}_j$ to denote the j-th column of U as a vector in $\mathbb{R}^M$. With this notation at hand we consider the following model for observed labels y:

$$y(j) = S(\mathbf{u}_j + \boldsymbol{\eta}_j), \qquad j \in Z', \tag{44}$$

where

$$\boldsymbol{\eta}_j = (\eta_{1j}, \cdots, \eta_{Mj})^T \in \mathbb{R}^M, \quad \text{and} \quad \eta_{mj} \overset{iid}{\sim} \psi.$$

Here ψ is a probability density function on $\mathbb{R}$ as before.

Remark 22 *Note that the assumption that η_{mj} are i.i.d. is not needed in general and one can consider correlations in the observation noise both between different classes and also amongst different points in the data set. However, for simplicity we only consider i.i.d. noise and leave the correlated noise setting for future study.* ◇

3.2. The One-Hot Likelihood

We begin by identifying the likelihood potential Φ for the model (44). For $j \in Z'$ and $m, \ell \in \{1, \cdots, M\}$ we have

$$\begin{aligned}
\mathbb{P}[y(j) = m|U] &= \mathbb{P}[u_{mj} + \eta_{mj} \geq u_{\ell j} + \eta_{\ell j}, \quad \forall \ell \in \{1, \cdots, M\}] \\
&= \mathbb{P}[\eta_{\ell j} \leq \eta_{mj} + u_{mj} - u_{\ell j}, \quad \forall \ell \in \{1, \cdots, M\}] \\
&= \mathbb{E}\Big[\mathbb{P}[\eta_{\ell j} \leq \eta_{mj} + u_{mj} - u_{\ell j}, \quad \forall \ell \in \{1, \cdots, M\}]\Big|\eta_{mj}\Big] \\
&= \mathbb{E}\Big[\prod_{\ell \neq m} \mathbb{P}[\eta_{\ell j} \leq \eta_{mj} + u_{mj} - u_{\ell j}]\Big|\eta_{mj}\Big] \\
&= \int_{\mathbb{R}} \psi(t) \prod_{\ell \neq m} \Psi(t + u_{mj} - u_{\ell j}) dt =: \breve{\Psi}(\mathbf{u}_j; m).
\end{aligned}$$

where Ψ is the CDF of ψ as in the binary case. To this end, we define the likelihood potential $\Phi(U; y)$ as

$$\Phi(U; y) = -\sum_{j \in Z'} \log \breve{\Psi}(\mathbf{u}_j; y(j)) = -\sum_{j \in Z'} \log\left(\int_{\mathbb{R}} \psi(t) \prod_{\ell \neq y(j)} \Psi(t + u_{y(j)j} - u_{\ell j}) dt\right), \tag{45}$$

which is in a similar form to (6).

3.3. Quadratic Regularization via Graph Laplacians (Multi-Class Case)

Recall, the matrix C_τ defined in (9) based on the graph Laplacian L. In a similar way to the probit method we define a quadratic regularization term for matrices $U \in \mathbb{R}^{M \times N}$ of the form

$$\langle C_\tau^{-1}, U^T U \rangle_F = \sum_{j,\ell=1}^{N} (C_\tau^{-1})_{j\ell} (U^T U)_{j\ell} = \sum_{m=1}^{M} \sum_{j,\ell=1}^{N} (C_\tau^{-1})_{j\ell} u_{m\ell} u_{mj}, \tag{46}$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. In the following we will us this quadratic term to regularize Problem 3 in the multi-class setting.

Remark 23 *If we think of C_τ^{-1} as a smoothing operator then the above choice for the regularization term promotes smoothness of the rows of U while the columns of U can be discontinuous. This means that each component of the function $u : Z \mapsto \mathbb{R}^M$ isomorphic to U is smooth amongst the vertices of G while the components themselves are allowed to be discontinuous at each node.* ◇

3.4. Properties of the One-Hot Minimizer

Putting together the one-hot likelihood in (45) and the quadratic regularization term (46) we define the *one-hot functional*

$$\mathcal{J}(U) := \frac{1}{2}\langle C_\tau^{-1}, U^T U\rangle_F + \Phi(U; y), \qquad U \in \mathbb{R}^{M\times N}. \tag{47}$$

We will see shortly that the one-hot functional has very similar properties to the probit functional J in binary classification. In particular, the regularization term (46) is strictly convex and provides stability and geometric information via the operator C_τ and the one-hot likelihood Φ is also convex and makes sure the minimizer of $\mathcal{J}$ is a good predictor of observed labels. We start by showing the convexity of the likelihood potential.

Proposition 24 (Convexity of the One-Hot Likelihood) *Let ψ be a log-concave PDF on $\mathbb{R}$. Then the function*

$$\breve{\Psi}(\mathbf{v}; m) = \int_{\mathbb{R}} \psi(t) \prod_{\ell\neq m} \Psi(t + v_m - v_\ell) dt, \qquad \mathbf{v} \in \mathbb{R}^M,$$

is log-concave on $\mathbb{R}^M$ for all $m \in \{1, \cdots, M\}$, and where Ψ is the CDF of ψ.

Proof By Bagnoli and Bergstrom (2005, Thm. 1) the functions Ψ are log-concave whenever ψ is log-concave. Furthermore, since log-concavity is preserved under affine transformations and finite products (see Saumard and Wellner, 2014, Sec 3.1) we conclude that $f(t, \mathbf{v}) = \psi(t)\prod_{\ell\neq m}\Psi(t + v_m - v_\ell)$ is log-concave on $\mathbb{R}^{M+1}$. The result now follows from the fact that the marginals of a log-concave function are also log-concave following Prékopa (1980, Thm. 3) and $\breve{\Psi}(\mathbf{v}; m)$ is precisely the marginal of $f(t, \mathbf{v})$ over the t variable. ■

Putting this result together with the fact that the quadratic term in (47) is strictly convex whenever C_τ^{-1} is strictly positive definite (which is true when $\tau^2 > 0$) gives the following result.

Proposition 25 *Let ψ be a continuous and log-concave PDF on $\mathbb{R}$ and let C_τ^{-1} be a strictly positive-definite matrix on $\mathbb{R}^N$. Then the functional $\mathcal{J}$ defined in (47) with Φ given by (45) is strictly convex.*

We are now set to prove an analog of Proposition 9 for the one-hot functional.

Proposition 26 (Representer Theorem for One-Hot Functional) *Let $G = \{X, W\}$ be a weighted graph and let ψ be a log-concave PDF. Suppose Φ is given by (45) and the matrix C_τ is given by (9) with parameters $\tau^2, \alpha > 0$. Then,*

(i) The one-hot functional $\mathcal{J}$ has a unique minimizer $U^* \in \mathbb{R}^{M\times N}$.

(ii) The minimizer U^* satisfies the EL equations

$$C_\tau^{-1} U^{*T} = \sum_{j\in Z'} \mathbf{e}_j (\mathbf{f}_j(\mathbf{u}_j^*))^T, \tag{48}$$

where $\mathbf{e}_j$ *is the* j*-th standard coordinate vector in* $\mathbb{R}^N$*, the vector* $\mathbf{u}_j^*$ *denotes the* j*-th column of* U^* *and the functions* $\mathbf{f}_j : \mathbb{R}^M \mapsto \mathbb{R}^M$ *are defined as*

$$\mathbf{f}_j(\mathbf{v}) = (f_{1j}(\mathbf{v}), \cdots, f_{Mj}(\mathbf{v}))^T, \qquad f_{mj}(\mathbf{v}) = \frac{1}{\breve{\Psi}(\mathbf{v}; y(j))} \frac{\partial \breve{\Psi}(\mathbf{v}; y(j))}{\partial v_m}, \tag{49}$$

for vectors $\mathbf{v} = (v_1, \cdots, v_M)^T$ *and*

$$\frac{\partial \breve{\Psi}(\mathbf{v}; m)}{\partial v_i} = \begin{cases} -\int_{\mathbb{R}} \psi(t)\psi(t + v_m - v_i) \prod_{\ell \neq i,m} \Psi(t + v_m - v_\ell) dt & \text{if} \quad i \neq m, \\ \sum_{k \neq m} \int_{\mathbb{R}} \psi(t)\psi(t + v_m - v_k) \prod_{\ell \neq k,m} \Psi(t + v_m - v_\ell) dt & \text{if} \quad i = m. \end{cases}$$

(iii) The minimizer U^* *can be represented using the expansion*

$$U^* = \sum_{j \in Z'} \breve{\mathbf{a}}_j \mathbf{c}_j^T,$$

where $\breve{\mathbf{a}}_j \in \mathbb{R}^M$ *and* $\mathbf{c}_j = C_\tau \mathbf{e}_j \in \mathbb{R}^N$*.*

(iv) The matrix U^* *solves* (48) *if and only if the vectors* $\breve{\mathbf{a}}_j = (\breve{a}_{1j}, \cdots, \breve{a}_{Mj})^T \in \mathbb{R}^M$ *solve the nonlinear system of equations*

$$\breve{\mathbf{a}}_j = \mathbf{f}_j \left(\sum_{k \in Z'} c_{jk} \breve{\mathbf{a}}_k \right), \qquad \forall j \in Z', \tag{50}$$

where c_{ij} *denote the entries of* C_τ*.*

Proof The method of proof is very similar to Proposition 9. (i) Follows directly from Proposition 25. (ii) Observe that $\mathcal{J}(U)$ is continuously differentiable, and so (48) follows by directly computing the first order optimality conditions $\nabla \mathcal{J}(U^*) = 0$. Proof of (iii) and (iv) is very similar to Proposition 9(iii) and (iv) and is essentially the result of solving the EL equations (48) directly. ■

Proposition 27 (One-Hot Dimension Reduction) *Suppose the conditions of Proposition 26 hold. Then*

(i) The problem of finding the matrix $U^* \in \mathbb{R}^{M \times N}$ *the minimizer of the one-hot functional* $\mathcal{J}$*, is equivalent to the problem of finding the Matrix* $B^* \in \mathbb{R}^{M \times J}$ *that solves*

$$(C'_\tau)^{-1} B^{*T} = \mathcal{F}'(B^*), \tag{51}$$

where C'_τ *is as in* (17) *and, for matrices* $B \in \mathbb{R}^{M \times J}$ *the map* $\mathcal{F}' : \mathbb{R}^{M \times J} \mapsto \mathbb{R}^{J \times M}$ *is defined by*

$$(\mathcal{F}'(B))_{im} := f_{m\pi^{-1}(i)}(\mathbf{b}_i) \quad \text{for} \quad (i, m) \in \{1, \cdots, J\} \times \{1, \cdots, M\},$$

where the reordering map π *is as in* (16) *and* $\mathbf{b}_i$ *denotes the* i*-th column of* B*.*

(ii) Moreover, the matrix B^* solves the optimization problem

$$B^* = \arg\min_{B\in\mathbb{R}^{J\times M}} \mathcal{J}'(B),$$

where

$$\mathcal{J}'(B) := \frac{1}{2}\langle (C'_\tau)^{-1}, B^T B\rangle_F - \sum_{i=1}^{J} \log \breve{\Psi}\Big(\mathbf{b}_i; y\left(\pi^{-1}(i)\right)\Big)$$

(iii) The matrices B^* and U^* satisfy the relationship

$$U^* = \sum_{j\in Z'} \breve{\mathbf{b}}_j^* \mathbf{c}_j^T,$$

where $\breve{\mathbf{b}}_j^*$ denotes the $\pi(j)$-th column of $B^*(C'_\tau)^{-T}$ and

$$\mathbf{b}_k^* = \mathbf{u}^*_{\pi^{-1}(k)}, \qquad k = \{1, \cdots, J\}.$$

Proof (i) Let $A = (a_{mi}) \in \mathbb{R}^{M\times J}$ be the matrix with entries $a_{mi} = \breve{\mathbf{a}}_{m\pi^{-1}(i)}$. That is, the columns of A are the $\breve{\mathbf{a}}_j$ vectors. Then we can rewrite (50) as

$$A = -\left(\mathcal{F}'(C'_\tau A^T)\right)^T, \tag{52}$$

Let

$$B^{*T} = C'_\tau A^T \in \mathbb{R}^{J\times M} \tag{53}$$

then we can rewrite (52) as

$$(C'_\tau)^{-1} B^{*T} = \mathcal{F}'(B^*).$$

(ii) Denote by b^*_{mi} the entries of B^*. Then we can directly verify that

$$(\mathcal{F}'(B^*))_{im} = -\frac{\partial}{\partial b^*_{mi}} \sum_{i=1}^{J} \log \breve{\Psi}\Big(\mathbf{b}_i^*; y(\pi^{-1}(i))\Big),$$

from which we infer that the matrix B^* indeed solves the following optimization problem

$$B^* = \arg\min_{B\in\mathbb{R}^{J\times M}} \frac{1}{2}\langle (C'_\tau)^{-1}, B^T B\rangle_F - \sum_{i=1}^{J} \log \breve{\Psi}\Big(\mathbf{b}_i; y(\pi^{-1}(j))\Big).$$

(iii) Following (53) $A = B^*(C'_\tau)^{-T}$. Let $\mathbf{a}_i$ denote the columns of A. Then by Proposition 26(iii),

$$U^* = \sum_{j\in Z'} \breve{\mathbf{a}}_j \mathbf{c}_j^T = \sum_{i=1}^{J} \breve{\mathbf{a}}_{\pi^{-1}(i)} \mathbf{c}^T_{\pi^{-1}(i)} = \sum_{i=1}^{J} \mathbf{a}_i \mathbf{c}^T_{\pi^{-1}(i)} = \sum_{j\in Z'} \breve{\mathbf{b}}_j^* \mathbf{c}_j^T.$$

On the other hand, using $B^* = A(C'_\tau)^T$ and Proposition 26(iii) we can write

$$b^*_{mk} = \sum_{i=1}^{J} a_{m\pi^{-1}(i)} c_{\pi^{-1}(i),\pi^{-1}(k)} = u_{m\pi^{-1}(k)},$$

which gives the desired identity connecting $\mathbf{b}_k^*$ and $\mathbf{u}^*_{\pi^{-1}(k)}$. ■

3.5. Consistency of the One-Hot Method

In analogy with binary classification we now discuss consistency of multi-class classification using the one-hot method. Our results here make use of the perturbation theory developed in Subsection 2.5. As in Subsection 2.6 we consider a graph $G_0 = \{X, W_0\}$ consisting of K connected clusters $\tilde{Z}_1, \cdots, \tilde{Z}_K$ and let $G_\epsilon = \{X, W_\epsilon\}$ be a family of graphs parameterized by $\epsilon > 0$ that are perturbations of G_0. We denote by $C_{\tau,\epsilon}$ the covariance matrix corresponding to the graph G_ϵ as defined in (30). Further, we assume the data y is generated by a ground truth function $U^\dagger \in \mathbb{R}^{M\times N}$ that is,

$$y(j) = S(\mathbf{u}_j^\dagger + \boldsymbol{\eta}_j), \qquad j \in Z', \tag{54}$$

where S is defined in (43) and we define the noise η_{mj} through a reference random variable analogous to (32):

$$\eta_{mj} = \gamma\breve{\eta}_{mj}, \qquad \breve{\eta}_{mj} \overset{iid}{\sim} \psi, \tag{55}$$

where ψ is a mean zero PDF with unit standard deviation. Similarly to Subsection 2.6, our first task is to estimate the probability of the event where the observed labels $y(j)$ are exact, i.e., $y(j)$ coincides with the index of the maximal element in the j-th column of $U^\dagger$.

Lemma 28 *Suppose ψ is log-concave then there exist constants $\omega_1, \omega_2 > 0$ so that*

$$\begin{aligned}&\mathbb{P}\big(y(j) = S(\mathbf{u}_j^\dagger),\ \forall j \in Z'\big)\\ &\quad\geq \int_{\mathbb{R}} \psi_\gamma(t) \prod_{j\in Z'} \left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}\left(t + \min_{k\neq y(j)}\left\{u^\dagger_{y(j)j} - u^\dagger_{kj}\right\}\right)\right)\right]^{M-1} dt.\end{aligned}$$

That is, if γ is small the data y is exact with high probability.

Proof Observe that $y(j) = S(\mathbf{u}_j^\dagger)$ whenever $u^\dagger_{y(j)j} + \eta_{y(j)j} \geq u^\dagger_{mj} + \eta_{mj}$ for all $m \neq y(j)$. Therefore,

$$\begin{aligned}\mathbb{P}\left(y(j) = S(\mathbf{u}_j^\dagger)\ \forall j \in Z'\right) &= \mathbb{P}(\{\eta_{mj} \leq \eta_{y(j)j} + u^\dagger_{y(j)j} - u^\dagger_{mj} : j \in Z', m \neq y(j)\})\\ &= \mathbb{E}\,\mathbb{P}\left(\{\eta_{mj} \leq \eta_{y(j)j} + u^\dagger_{y(j)j} - u^\dagger_{mj}, : j \in Z', m \neq y(j)\}|\eta_{y(j)j}\right).\end{aligned}$$

Now by Lemma 15 and Markov's inequality (see proof of Lemma 16) along with independence of the η_{mj} we conclude there exist constants $\omega_1, \omega_2 > 0$ so that for fixed $j \in Z'$,

$$\begin{aligned}&\mathbb{P}\big(\{\eta_{mj} \leq \eta_{y(j)j} + u^\dagger_{y(j)j} - u^\dagger_{mj}, : m \neq y(j)\}|\eta_{y(j)j}\big)\\ &\quad\geq \mathbb{P}\big(\{\eta_{mj} \leq \eta_{y(j)j} + \min_{k\neq y(j)}\{u^\dagger_{y(j)j} - u^\dagger_{kj}\} : m \neq y(j)\}|\eta_{y(j)j}\big)\\ &\quad\geq \prod_{m\neq y(j)}\left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}\left(\eta_{y(j)j} + \min_{k\neq y(j)}\{u^\dagger_{y(j)j} - u^\dagger_{kj}\}\right)\right)\right]\\ &\quad= \left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}\left(\eta_{y(j)j} + \min_{k\neq y(j)}\{u^\dagger_{y(j)j} - u^\dagger_{kj}\}\right)\right)\right]^{M-1}.\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\{\eta_{mj} \le \eta_{y(j)j} + u^\dagger_{y(j)j} - u^\dagger_{mj}, : j \in Z', m \ne y(j)\} | \eta_{y(j)j}\right)$$
$$\ge \prod_{j\in Z'}\left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}\left(\eta_{y(j)j} + \min_{k\ne y(j)}\left\{u^\dagger_{y(j)j} - u^\dagger_{kj}\right\}\right)\right)\right]^{M-1}.$$

Integrating the above bound over $\eta_{y(j)j}$ gives the desired result. ■

The following lemma is the analog of Lemma 17 for the one-hot model (54). The method of proof is identical to that lemma and is therefore omitted.

Lemma 29 *Suppose* (54) *and* (55) *hold,* ψ *is log-concave and*

$$\min_{m,k\in\{1,\cdots,M\}, m\ne k}\{|u^\dagger_{mj} - u^\dagger_{kj}|\} > \theta > 0 \qquad \forall j \in Z'. \tag{56}$$

Then for any sequence $\gamma \downarrow 0$, $y(j) \xrightarrow{\text{a.s.}} S(\mathbf{u}^\dagger_j)$ *with respect to* $\prod_{(m,j)\in\{1,\ldots,M\}\times Z'} \psi(t_{mj})$ *the law of the i.i.d. sequence* $\{\breve{\eta}_{mj}\}_{(m,j)\in\{1,\cdots,M\}\times Z'}$.

With the above lemmata at hand we are now in a position to study consistency of minimizers of the one-hot functional

$$\mathcal{J}_{\tau,\epsilon,\gamma}(U) := \frac{1}{2}\langle C^{-1}_{\tau,\epsilon}, U^T U\rangle_F + \Phi_\gamma(U; y), \qquad U \in \mathbb{R}^{M\times N}, \tag{57}$$

where

$$\Phi_\gamma(U; y) := -\sum_{j\in Z'} \log \breve{\Psi}_\gamma(\mathbf{u}_j; y(j)), \qquad \breve{\Psi}_\gamma(\mathbf{v}; m) := \int_{\mathbb{R}} \psi_\gamma(t) \prod_{\ell\ne m} \Psi_\gamma(t + v_m - v_\ell) dt$$

and Ψ_γ is the CDF of $\psi_\gamma(\cdot) := \frac{1}{\gamma}\psi(\frac{\cdot}{\gamma})$ as before.

3.5.1. One-Hot Consistency With A Single Observed Label

Once again we start with the case of a single observed label with $Z' = \{1\}$ belonging to the first cluster $\tilde{Z}_1$ and without loss of generality we assume $u^\dagger_{11} > u^\dagger_{m1}$ for all $m \ne 1$ so that the correct label of the observed node is 1.

Proposition 30 *Consider the single observation setting above and suppose Assumptions 1 and 2 are satisfied by* G_0 *and* G_ϵ. *Let* U^* *be the minimizer of* $\mathcal{J}_{\tau,\epsilon,\gamma}$ *and let* $\gamma > 0$.

(a) If $\epsilon = o(\tau^2)$ *as* $\tau \to 0$ *then* $\exists \tau_0 > 0$ *so that* $\forall (\tau,\gamma) \in (0,\tau_0)\times(0,\infty)$ *and* $\forall j \in \tilde{Z}_1$

$$\mathbb{P}\Big(S(\mathbf{u}^*_j) = 1\Big) \ge \int_{\mathbb{R}} \psi_\gamma(t)\left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}\left(t + \min_{m\ne 1}\left\{u^\dagger_{11} - u^\dagger_{m1}\right\}\right)\right)\right]^{M-1} dt, \tag{58}$$

where $\omega_1, \omega_2 > 0$ *are uniform constants depending only on* ψ.

(b) If $\epsilon = \Theta(\tau^2)$ *then the above statement holds for all* $j \in Z$.

Proof (a) The method of proof is very similar to that of Proposition 18. The dimension reduced system (51) takes the simpler form

$$\mathbf{b}_1^* = (C_{\tau,\epsilon})_{11}\mathbf{f}_1(\mathbf{b}_1^*). \tag{59}$$

As before, if $\epsilon = o(\tau^2)$ by Proposition 14(a) there exists $\tau_0 > 0$ so that $\forall \tau \in (0, \tau_0)$ (39) holds and so, up to leading order (59) is equivalent to

$$\mathbf{b}_1^* = |\,(\bar{\boldsymbol{\chi}}_1)_1\,|^2\mathbf{f}_1(\mathbf{b}_1^*).$$

Now consider the event where $y(1) = S(\mathbf{u}_1^\dagger) = 1$, i.e., the data is exact. Then, we immediately see from (49) and the fact that ψ_γ and Ψ_γ are positive that, the only entry of $\mathbf{f}_1(\mathbf{b}_1^*)$ that is not negative is the first entry and so $b_{11}^* > 0$ while $b_{m1}^* < 0$ for all $m \neq 1$. Finally, by Proposition 27(iii) we can write

$$U^{*T} = \mathbf{c}_{1,\epsilon} \cdot \mathbf{f}_1(\mathbf{b}_1^*)^T = (\bar{\boldsymbol{\chi}}_1)_1\, \bar{\boldsymbol{\chi}}_1 \cdot \mathbf{f}_1(\mathbf{b}_1^*)^T + \mathcal{O}\left(\epsilon/\tau^2 + \tau^{2\alpha} + \epsilon\right).$$

It is then straightforward to see that when τ is sufficiently small then for all $j \in \tilde{Z}_1$ we have $S(\mathbf{u}_j^*) = y(1) = S(\mathbf{u}_1^\dagger) = 1$ and the claim follows by bounding the probability of the event where $y(1) = 1$ using Lemma 28. Part (b) follows by a very similar argument to proof of Proposition 18(b). ■

The following corollary is the analogue of Corollary 19 for the one-hot method. The proof follows from the proof of Proposition 30 and Lemma 29.

Corollary 31 *Suppose Proposition 30 and condition* (56) *are satisfied. Then the following holds with a.s. convergence in the sense of Lemma 29:*

(a) For any sequence $\gamma, \tau, \epsilon \downarrow 0$ along which $\epsilon = o(\tau^2)$ it holds that

$$S\left(\mathbf{u}_j^*\right) \xrightarrow{\text{a.s.}} S\left(\mathbf{u}_1^\dagger\right), \qquad \forall j \in \tilde{Z}_1.$$

(b) For any sequence $\gamma, \tau, \epsilon \downarrow 0$ along which $\epsilon = \Theta(\tau^2)$ the above statement holds for all $j \in Z$.

3.5.2. One-Hot Consistency With Multiple Observed Labels

We finally consider the general setting where multiple labels are observed and $|Z'| = J \geq 2$. We need the analog of Assumption 3 in multi-class classification:

Assumption 4 *Let $\tilde{Z}_k$ be a cluster within which a label has been observed, i.e., $\tilde{Z}_k \cap Z' \neq \varnothing$. Then $S(\mathbf{u}_j^\dagger)$ is constant for all $j \in \tilde{Z}_k$.*

Proposition 32 *Consider the multiple observation setting above and suppose Assumptions 1, 2 and 4 are satisfied by G_0, G_ϵ and $U^\dagger$. Let U^* be the minimizer of $\mathcal{J}_{\tau,\epsilon,\gamma}$ and Z'' be as in* (40) *the set of nodes in clusters for which labels have been observed. If $\epsilon = o(\tau^2)$ as $\tau \to 0$ then $\exists \tau_0 > 0$ so that $\forall (\tau, \gamma) \in (0, \tau_0) \times (0, \infty)$ and $\forall j \in Z''$*

$$\mathbb{P}\Big(S\left(\mathbf{u}_j^*\right) = S\left(\mathbf{u}_j^\dagger\right)\Big) \geq \int_{\mathbb{R}} \psi_\gamma(t) \prod_{j \in Z'} \left[1 - \omega_2 \exp\left(-\frac{\omega_1}{\gamma}\left(t + \min_{m \neq y(j)}\left\{u_{y(j)j}^\dagger - u_{mj}^\dagger\right\}\right)\right)\right]^{M-1} dt,$$

where $\omega_1, \omega_2 > 0$ are uniform constants depending only on ψ.

Proof The proof follows similar steps to the binary result in Proposition 20 and we use the same notation as in the proof of that result. Here the dimension reduced system (51) takes the form

$$(C'_{\tau,\epsilon})^{-1} B^{*T} = \mathcal{F}'(B^*). \tag{60}$$

Then, Proposition 41 implies that $C'_{\tau,\epsilon}$ is nearly block diagonal and (42) holds. Then, up to leading order (60) takes the form

$$(\bar{\boldsymbol{\chi}}_k)_j^{-1} \mathbf{b}^*_{\pi(j)} = \sum_{i \in \tilde{Z}'_k} (\bar{\boldsymbol{\chi}}_k)_i \mathbf{f}_i(\mathbf{b}^*_{\pi(i)}), \qquad \text{for } j \in \tilde{Z}'_k \text{ and } k \in \{1, \cdots, K\},$$

where we used the same notation as in (49) and made use of the approximation (42) for the elements of $C'_{\tau,\epsilon}$. Once again the right hand side of the above expression is independent of j and so the $(\bar{\boldsymbol{\chi}}_k)_j^{-1} \mathbf{b}^*_{\pi(j)}$ vectors are constant for all $j \in \tilde{Z}'_k$. We then have

$$\mathbf{b}^*_{\pi(j)} = (\bar{\boldsymbol{\chi}}_k)_j \sum_{i \in \tilde{Z}'_k} (\bar{\boldsymbol{\chi}}_k)_i \mathbf{f}_i \left(\frac{(\bar{\boldsymbol{\chi}}_k)_i}{(\bar{\boldsymbol{\chi}}_k)_j} \mathbf{b}^*_{\pi(j)} \right), \qquad \text{for } j \in \tilde{Z}'_k \text{ and } k \in \{1, \cdots, K\}.$$

Now consider the event where the data y is exact which happens with high probability following Lemma 28. Since $U^\dagger$ satisfies Assumption 4 then $y(j)$ is constant for all $j \in \tilde{Z}'_k$. Using the definition of $\mathbf{f}_j$ given in (49), we directly verify that the only positive coordinate of $\mathbf{f}_i \left(\frac{(\bar{\boldsymbol{\chi}}_k)_i}{(\bar{\boldsymbol{\chi}}_k)_j} \mathbf{b}^*_{\pi(j)} \right)$ is the $y(j)$-th coordinate while all other coordinates are negative. Since $(\bar{\boldsymbol{\chi}}_k)_i$ and $(\bar{\boldsymbol{\chi}}_k)_j$ are both strictly positive in the above, we conclude $S(\mathbf{b}^*_{\pi(j)}) = y(j) = S(\mathbf{u}^\dagger_j)$. The desired result now follows by Proposition 27(iii). ■

Similar to Corollary 31 we obtain the following result by applying Theorem 32 and Lemma 29.

Corollary 33 *Suppose Theorem 32 and condition* (56) *are satisfied. Then for any sequence* $\gamma, \tau, \epsilon \downarrow 0$ *along which* $\epsilon = o(\tau^2)$ *it holds that*

$$S\left(\mathbf{u}^*_j\right) \xrightarrow{\text{a.s.}} S\left(\mathbf{u}^\dagger_j\right), \qquad \forall j \in Z'',$$

with a.s. convergence in the sense of Lemma 29.

4. Numerical Experiments

In this section we turn our attention to numerical experiments that are designed to confirm our theoretical findings, and to expand upon the behavior of the probit and one-hot methods beyond our theory. In Subsection 4.1 we study the spectrum of graph Laplacians on a graph consisting of three clusters that are weakly connected reaffirming the analysis of Appendix A. Subsection 4.2 is dedicated to the consistency of the probit method. Here we demonstrate a curve in the $(\epsilon/\tau^2, \alpha)$ plane across which probit transitions from being consistent to propagating the majority label. Finally, we repeat similar experiments for the one-hot method in Subsection 4.3 demonstrating a similar phase transition curve in the $(\epsilon/\tau^2, \alpha)$ plane and showing that the curve is sensitive to the number of observed labels in different clusters.

4.1. Spectrum of Covariance Operators

We begin with a numerical demonstration of the perturbation theory of Appendix A and in particular the result of Proposition 39. At the same time we will introduce a synthetic experiment which is used throughout this section for illustration.

We consider a random data set using points drawn from a mixture of three Gaussian distributions centered at $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$ with variance 0.1. We draw $N = 150$ points in total and the data set has three clusters with 50 points in each one. In order to construct a weighted graph on this data set we choose the kernel function

$$\kappa(t) = \mathbf{1}_{\{t \le 0.25\}}(t),$$

and set $w_{ij}^{(0)} = \kappa(|x_i - x_j|)$ where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^3$. Figure 1(a) shows the data points as well as the connected components of the resulting weighted graph in this example. With the weight matrix W_0 at hand we define the graph Laplacian operator $L_0 = D_0 - W_0$ and let $C_{\tau,0}^{-1} = \tau^{-2\alpha}(L_0 + \tau^2 I)^\alpha$. In other words, we fix $p = 0$ and so the functions $\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$ simplify to the indicator functions on the clusters $\tilde{Z}_1, ..., \tilde{Z}_K$. We further define a perturbation of this matrix by replacing W_0 with W_ϵ with entries $w_{ij}^{(\epsilon)} = \kappa_\epsilon(|x_i - x_j|)$ where the perturbed kernel has the form

$$\kappa_\epsilon(t) = \kappa(t) + \epsilon \exp\left(-\frac{t^2}{(0.25+\epsilon)^2}\right). \tag{61}$$

Clearly, the resulting weighted graph for any positive value of ϵ is fully connected but the weight of edges connecting the three clusters are at most of order ϵ. We consider the perturbed covariance matrix $C_{\tau,\epsilon}^{-1} = \tau^{-2\alpha}(L_\epsilon + \tau^2 I)^\alpha$ and study its low-lying spectrum. In the notation of Appendix A we use $\lambda_{2,\epsilon}$ and $\lambda_{3,\epsilon}$ to denote the first two non-trivial eigenvalues of $C_{\tau,\epsilon}$. Figure 1(b) shows that as ϵ/τ^2 becomes small $\lambda_{j,\epsilon} - 1$ vanishes linearly in ϵ/τ^2 for $j = 2, 3$ which is in perfect agreement with Proposition 39(ii). We also observe that $\lambda_{4,\epsilon}$ blows up with τ^{-2} as predicted in Proposition 40. On the other hand, Figure 1(c) shows the ℓ^2 distance between the second and third eigenvectors of $C_{\epsilon,\tau}$ with their projection onto the span of the set functions $\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^3$. We displayed the case $\alpha = 1$, the behavior for other choices of α is similar. Here P_0 denotes the projection onto the span of $\{\boldsymbol{\chi}_k\}_{k=1}^3$ following the notation of Appendix A. We observed that $\|(I - P_0)\boldsymbol{\phi}_{j,\epsilon}\|_2$ goes to zero linearly in ϵ for $j = 2, 3$, which is precisely the rate predicted in Proposition 39(iii) suggesting that the bound is sharp.

4.2. Binary Classification With Probit

We now consider a binary classification problem on the synthetic data set of Subsection 4.1. Figure 2(a) shows the true label of the 150 points in the data set. Blue points have label +1 while red points have label −1. For the SSL problem we assume one label is observed within each cluster and that the observed labels are correct, i.e., $y = (+1, +1, -1)^T$. We take the observation noise η_j to be i.i.d. logistic random variables with mean zero. More precisely,

$$\psi_\gamma(t) = \frac{\exp(-t/\gamma)}{\gamma(1+\exp(-t/\gamma))}, \qquad \Psi_\gamma(t) = (1+\exp(-t/\gamma))^{-1}.$$

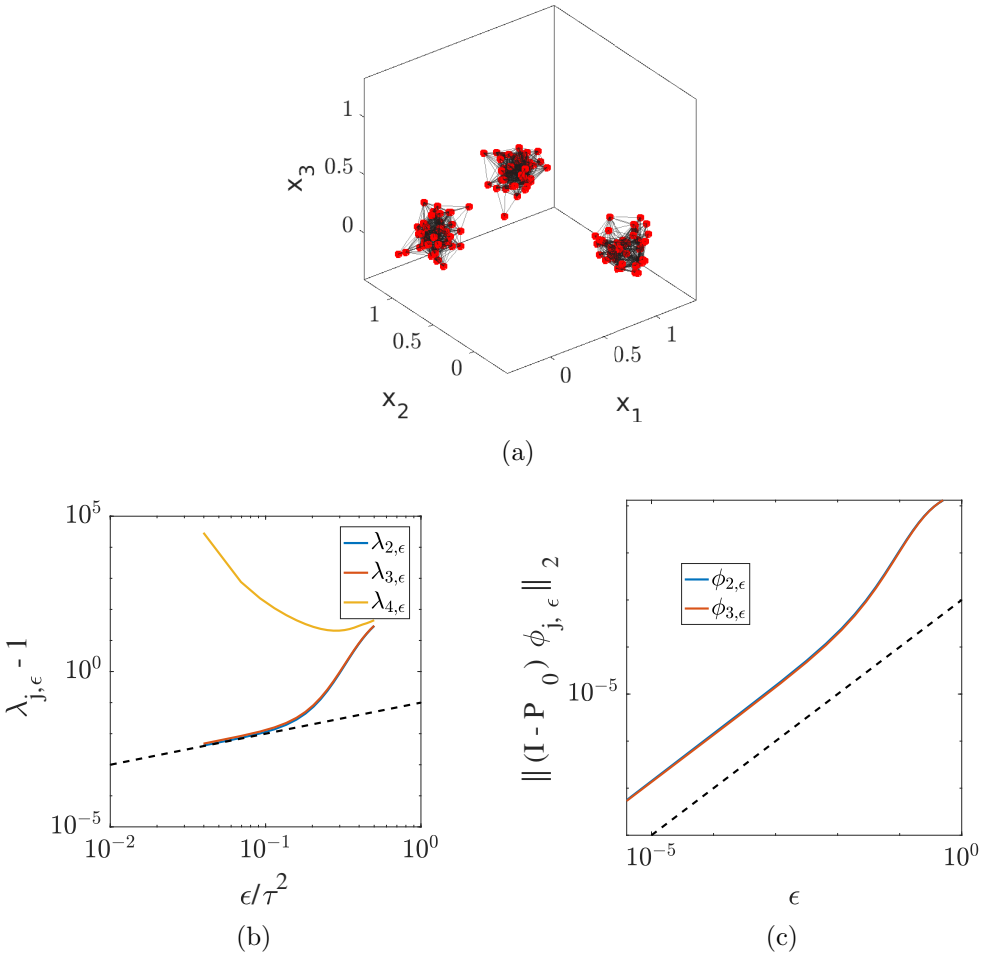


Figure 1: (a) The disconnected graph G_0 constructed by random draws from a mixture of three Gaussians. (b) The first three eigenvalues of the perturbed inverse covariance operator $C_{\tau,\epsilon}^{-1}$ with $\alpha = 1$ for different choices of ϵ and τ^2. The fourth eigenvector $\lambda_{4,\epsilon}$ blows up with τ^{-2} as predicted while $\lambda_{2,\epsilon}$ and $\lambda_{3,\epsilon}$ converge to 1 linearly in ϵ/τ^2. (c) The distance between $\boldsymbol{\phi}_{2,\epsilon}$ and $\boldsymbol{\phi}_{3,\epsilon}$ and their projections onto the span of the set functions $\{\boldsymbol{\chi}_j\}_{j=1}^3$. The distance vanishes linearly in ϵ.

We use the κ_ϵ kernel of (61) and construct the graph Laplacian L_ϵ and the covariance operator $C_{\tau,\epsilon}$ as in subsection 4.1 above. Since the matrix $C_{\epsilon,\tau}^{-1}$ is ill-conditioned for small τ and large values of α we found it crucial to use a low-rank approximation to $C_{\epsilon,\tau}$ in order to solve the EL equations (12) in a stable manner. Following the expansion (21) and the fact that there exists a uniform spectral gap between $\lambda_{3,\epsilon}$ and $\lambda_{4,\epsilon}$ (recall Figure 1(b)) we consider the truncated expansion

$$\hat{C}_{\epsilon,\tau} = \sum_{k=1}^{n} \frac{1}{\lambda_{k,\epsilon}} \boldsymbol{\phi}_{k,\epsilon} \boldsymbol{\phi}_{k,\epsilon}^T, \tag{62}$$

where a suitable truncation $n < N$ will be chosen later, and solve the approximate EL equations

$$\mathbf{u}^* = \sum_{j \in Z'} \hat{C}_{\epsilon,\tau} F_j(u_j^*) \mathbf{e}_j.$$

We used MATLAB's `fsolve` function for this task. For our first set of experiments we fix the noise parameter $\gamma = 0.5$ and used $n = 10$ terms in the approximation of $\hat{C}_{\tau,\epsilon}$. We then vary ϵ, τ and α. We considered $\tau \in (0.01, 1)$, $\epsilon/\tau^2 \in (0.01, 0.5)$, and $\alpha \in (0.25, 10)$, i.e., for each value of α we pick two sequences of τ and ϵ/τ^2 values and set $\epsilon = \tau^2 \times \epsilon/\tau^2$. Figure 2(b) shows the percentage of mislabelled points when $\operatorname{sgn}(\mathbf{u}^*)$ is used as the label predictor. The maximum error of 33% corresponds to all red points being labelled as blue. Our results suggest that when ϵ/τ^2 is large the Probit classifier $\mathbf{u}^*$ tends to assign the majority labels to all points in the data set. Furthermore, we observe a sharp transition between perfect label recovery and assignment of majority labels. This effect is amplified for larger values of α in that the transition seems to happen for a smaller value of ϵ/τ^2.

We also consider the distance between the span of the set functions $\boldsymbol{\chi}_j$ and the minimizer $\mathbf{u}^*$. Figure 2(c) shows $\|(I - P_0)\mathbf{u}^*\|_2$ as a function of τ^2 and for different values of α. We see that for smaller values of α the projection error is $\mathcal{O}(\tau^{2\alpha})$ which is in line with the predicted error between $\mathbf{c}_{j,\epsilon}$ and the span of $\boldsymbol{\chi}_j$ in Proposition 14(a) since for smaller values of α the leading order error term is $\mathcal{O}(\tau^{2\alpha})$. When α is large the projection error is controlled by the $\mathcal{O}(\epsilon/\tau^2 + \epsilon)$ terms in the error bound of Proposition 14(a) and no longer depends on α.

We noticed that the labelling accuracy is more or less independent of γ so long as the data y is correct as demonstrated in Figure 3. We also note that, for the most part, the labelling accuracy is independent of n as well. As shown in Figure 3 the labelling accuracy increases slightly only for small values of α and larger values of ϵ/τ^2.

For our final set of experiments we consider noisy data. First, we fix $\tau = 0.5$, $\epsilon/\tau^2 = 0.1$, $n = 10$, and take $\alpha \in (0.25, 10)$ and $\gamma \in (0.1, 1)$. For each value of α and γ we perform 100 experiments where we randomly perturb the data y by drawing independent measurement noise η_j using the model (31) and consider the labelling accuracy of the probit model. If all labels are recovered correctly by $\operatorname{sgn}(\mathbf{u}^*)$ we consider the experiment a success and otherwise a failure.

In Figure 4(a), we plot the probability of success of predicting the correct label of all points as a function of γ and ϵ/τ^2 for fixed $\alpha = 2$. We chose $\tau = 0.5$, $\epsilon/\tau^2 \in (0.1, 0.6)$ and $\gamma \in (0.1, 1)$. Here we see a clear transition in the success probability as a function of ϵ/τ^2. When ϵ/τ^2 is small the success probability is almost independent of ϵ/τ^2 and depends only on γ but for larger values of ϵ/τ^2 the success probability suddenly drops to zero meaning that some points are always mislabelled. This behavior is in line with Figure 3 where we

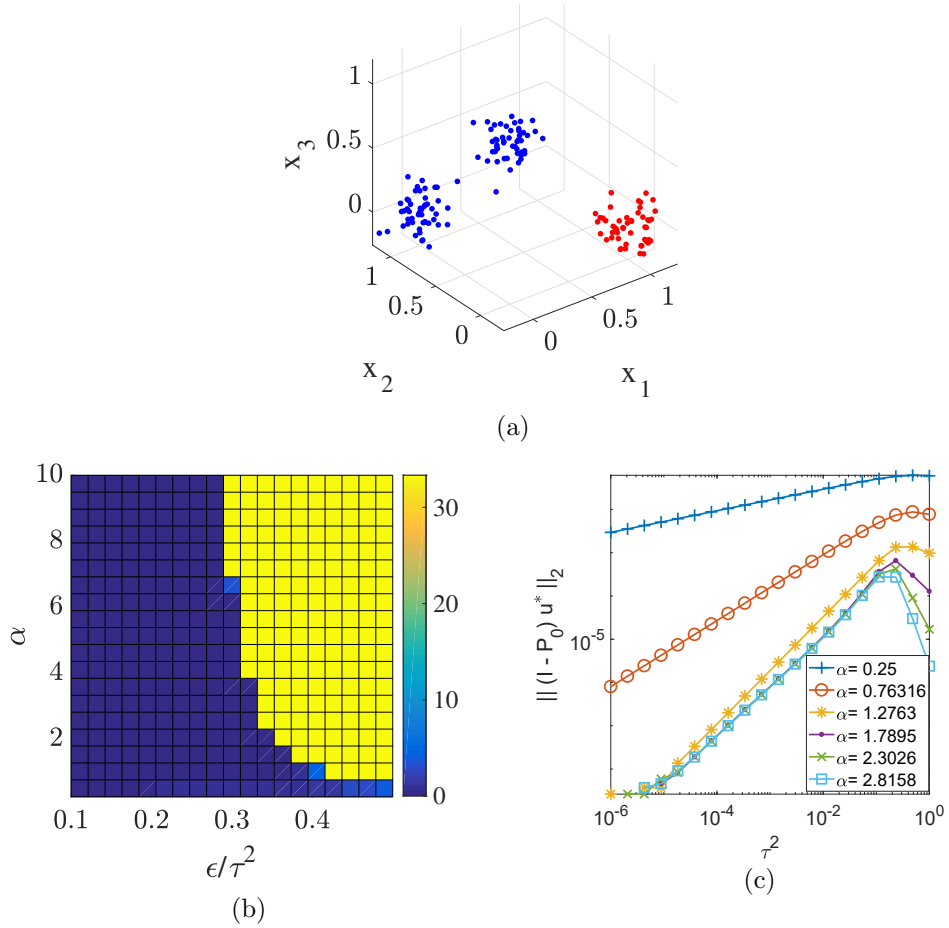


Figure 2: (a) The true binary labels in the synthetic data set of subsection 4.2. Blue points have label +1 and red points have label −1. (b) Heat map of the percentage of mislabelled points using the probit classifier. The 33% error mark corresponds to assigning label +1 to all red points which constitute a third of the data set. For fixed values of α we observe a sharp transition as ϵ/τ^2 increases where we go from labelling all of the points correctly to labelling the red points as blue. (c) The distance between $\mathbf{u}^*$ and the span of $\{\boldsymbol{\chi}_j\}_{j=1}^3$. This distance is $\mathcal{O}(\tau^{2\alpha})$ when α is small and does not depend on α when it is large indicating that the distance is controlled by $\mathcal{O}(\epsilon/\tau^2 + \epsilon)$.

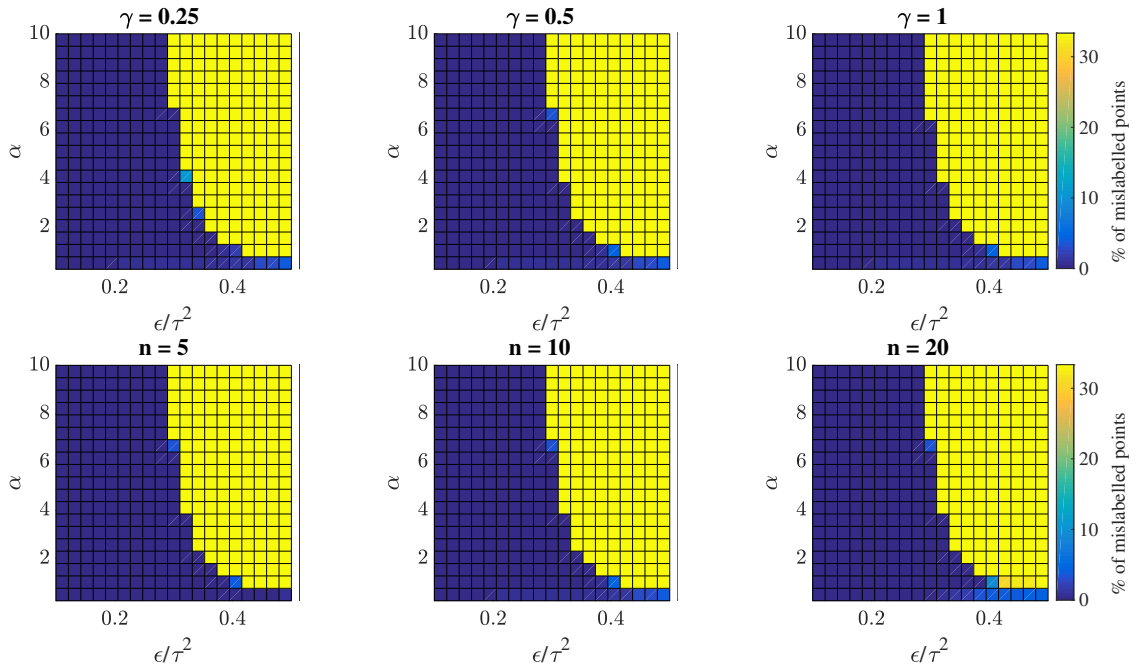


Figure 3: Demonstrating the labelling accuracy of probit as a function of measurement noise γ and the truncation parameter n. In the top row we fix $n = 10$ and modify γ, while in the bottom row we fix $\gamma = 0.5$ and modify n. We do not observe much sensitivity to n, in particular when α is large. When α is small we can see a slight increase in error for large values of ϵ/τ^2 as n grows larger. Modifying γ does not have a significant impact so long as the data y is correct.

observed a sharp increase in the prediction error when ϵ/τ^2 is too large. We emphasize that this behavior is also in line with Proposition 20 stating that the probability of success is controlled only by γ provided that ϵ/τ^2 and ϵ are sufficiently small.

Next, we fixed $\epsilon/\tau^2 = 0.1$ and modified α, see Figure 4(b). We do not observe any dependence of the success probability on α. This is in line with Proposition 20, which states that if ϵ and ϵ/τ^2 are sufficiently small then the success probability is essentially controlled by the probability of the event where the data is correct which depends only on γ.

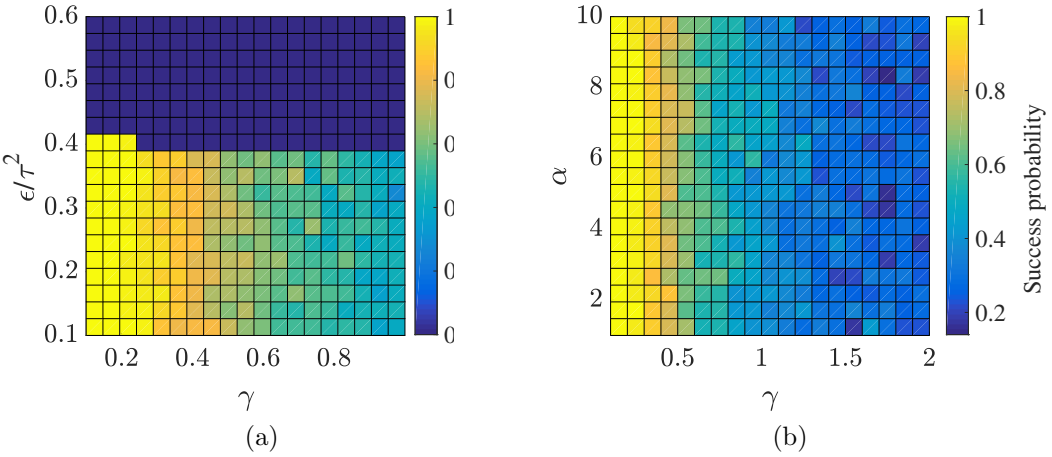


Figure 4: Heat map of the probability of success of probit in predicting the correct label of all the points in the data set. (a) We plot the success probability of probit as a function of ϵ/τ^2 and γ for fixed value of $\alpha = 2$. When ϵ/τ^2 is small the success probability appears to only depend on γ but as ϵ/τ^2 increases we see a sharp transition where the success probability drops to zero indicating that some points are always mislabelled in this regime. (b) The success probability for different values of α and γ for fixed value of $\epsilon/\tau^2 = 0.1$. We do not observe any strong dependence on α and the success probability appears to depend on γ only.

4.3. Multi-Class Classification Using One-Hot

For the next set of numerical experiments we consider multi-class classification with the One-hot method. Once again we use the synthetic data set of subsection 4.1 but now we assume there exist three classes within the graph as depicted in Figure 5. Similar to previous sections we take ψ_γ to be the logistic distribution but this time use the model (44) for the observed labels. We are assuming that we are given one label in each cluster, i.e. $J = 3$. In the perfect measurement case $y = (1, 2, 3)^T$. The main modification in this case, as compared to binary classification is that now we need to minimize the one-hot functional which has a more complicated misfit function as in (45). Furthermore, the minimizer is now a matrix $U^* \in \mathbb{R}^{M\times N}$ where $M = 3$ and $N = 150$. In this case (48) is a nonlinear system of equations in 3×150 dimensions which is slow to solve. Instead, we solve the dimension reduced system

(51) and identify U^* via B^*. We noticed that this approach offers significant speedup in our calculations. The dimension reduced system (51) is small enough that MATLAB's `fsolve` can still be very effective. We highlight that we approximate the matrix $C'_{\epsilon,\tau}$ by finding $\hat{C}_{\epsilon,\tau}$ as in (62) and then keep only the rows and columns of $\hat{C}_{\epsilon,\tau}$ that correspond to the observation vertices in Z'.

Overall we find that the one-hot method behaves similarly to probit as expected following our analysis. In Figures 6 we show the accuracy of the one-hot method in predicting the correct label of the points using perfect observed labels y. Similarly to the probit case we see little sensitivity to n and γ but a clear phase transition in the prediction error as ϵ/τ^2 increases for each value of α. We see that the prediction error is either very small or close to 66%. The latter value is a result of all three clusters being labelled as the same class. Overall, it seems that the prediction error is smaller and less sensitive to ϵ/τ^2 when α is small.

Similarly to the probit case we also study the success probability of one-hot for different values of α, ϵ/τ^2 and γ. Here we say that the one-hot classification is successful if all labels within the data set are predicted correctly. As before we vary the value of γ between 0 and 1 and estimate the success probability of one-hot by averaging over 100 trial runs with randomly perturbed observed labels y. Figure 7(a) and (b) show a heat-map of success probability of one-hot for different values of ϵ/τ^2 and γ. Here we observe some dependence between the success probability and ϵ/τ^2. In fact, for fixed value of γ we see a small increase in success probability as ϵ/τ^2 increases. However, as depicted in Figure 7(b) increasing ϵ/τ^2 eventually leads to a sudden drop in probability of success. This behavior is again in line with the phase transition observed in Figure 6. On the other hand, we observe in Figure 7(c) that for fixed values of ϵ/τ^2 the success probability is effectively independent of α and only controlled by γ. This is in line with our numerical results in the binary case, see Figure 4(a).

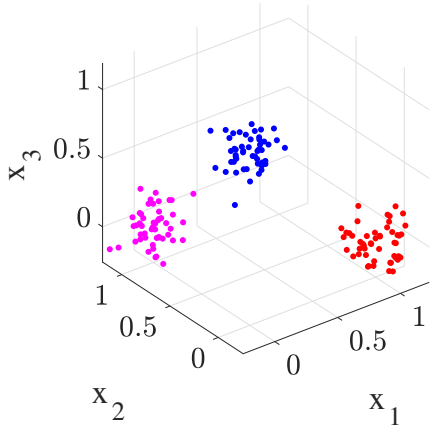


Figure 5: Visualization of the three classes within the three cluster data set for multi-class classification using one-hot. We observe a single label in each cluster.

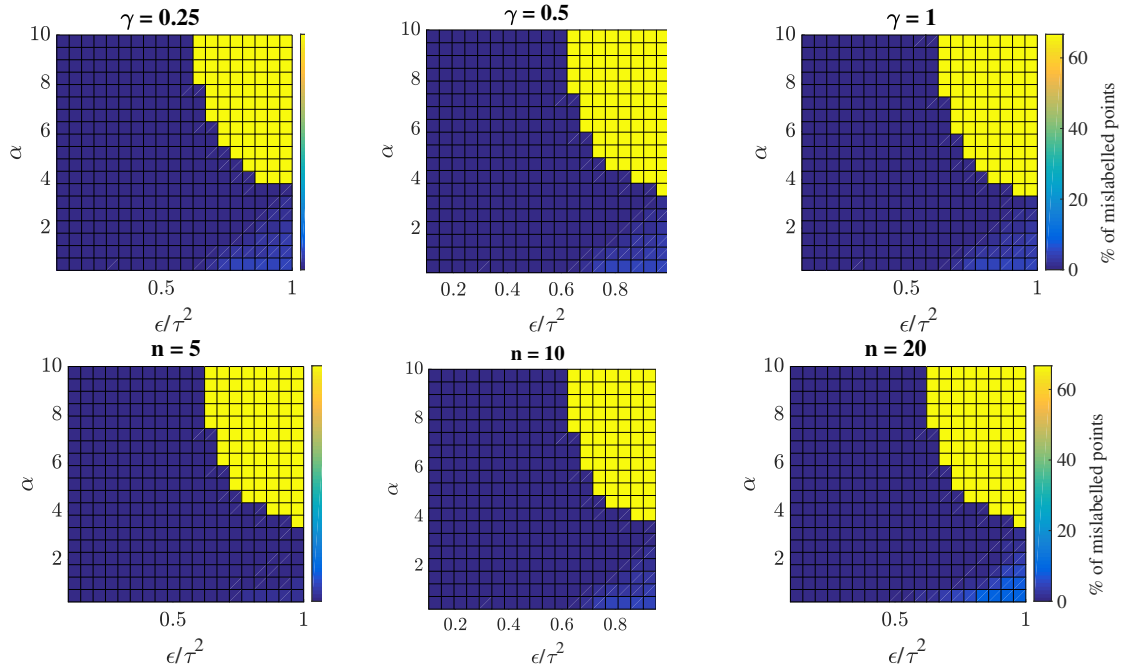
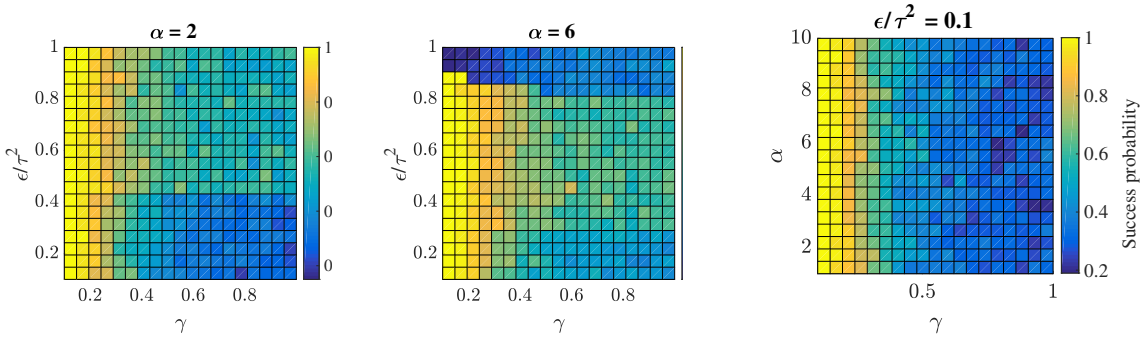


Figure 6: Prediction error of the one-hot classifier using perfect measurements for different values of γ while fixing $n = 10$ (top row) and values of n while fixing $\gamma = 0.5$ (bottom row). A clear phase transition is observed as ϵ/τ^2 increases for each value of α. For large values of ϵ/τ^2 and small α we observe a slight increase in prediction error as n increases (compare to bottom row when $\alpha \approx 1$).



(a) (b) (c)

Figure 7: Heat map of success probability of one-hot as a function of γ, α and ϵ/τ^2. (a) and (b) show the success probability as a function of ϵ/τ^2 for $\alpha = 2$ and $\alpha = 6$. In the latter case we see a sudden drop in the success probability as ϵ/τ^2 increases, even for very small values of γ. (c) shows the success probability as a function of α and γ for fixed $\epsilon/\tau^2 = 0.1$. Here the success probability is mostly controlled by γ and does not appear to depend on α.

4.3.1. Effect of Number of Observations In Each Cluster

For our final set of observations we consider varying the number of observations in each cluster. We use the same data set as above with $N = 150$ in three distinct clusters but this time we vary the number of observed labels in each cluster. In Figure 8 (a) and (b) we compare the accuracy of one-hot with a single observation in all clusters and with three observations in all clusters respectively. We only consider noiseless observations in this case. We see a small perturbation in the error phase transition for larger number of observations but overall the accuracy appears to depend weakly on the number of observations so long as we have the same number of observations in all clusters. Figure 8(c) shows the same experiment as above except that here we took two observations in one of the clusters and a single observation in the other two. Figure 9(d) shows a similar calculation with three observations in one cluster and a single observation in the rest. In comparison to Figure 8(a) and (b) we now see a major change in the accuracy of one-hot, namely that the jump in error now occurs for smaller values of ϵ/τ^2 indicating that the majority label of one of the clusters is propagated to the rest of the data set when we have an unbalanced number of observations within the clusters.

In Figure 9 we show the success probability of one-hot when an unbalanced number of labels are observed in the clusters: three labels are observed in one cluster and single labels are seen in the rest. As in previous success probability calculations we randomly perturbed the observed labels in 100 trials and approximated the success probability of one-hot as a function of ϵ/τ^2 and γ. As before, we see a sharp transition in the success probability as ϵ/τ^2 grows but below a certain critical value of ϵ/τ^2 the success probability appears to only depend on γ. We also note that this critical value of ϵ/τ^2 appears to shift towards smaller values for larger α.

These experiments reveal an interesting and complicated feature of the probit and one-hot minimizers in connection to the balancing of labelled points in clusters that warrant future analysis. It appears that having a balanced number of labels in different clusters allows for a larger range of acceptable ϵ, τ^2 parameters; note that the blue regions are larger in Figure 8(a,b) when all clusters have the same number of labelled points compared to Figure 8(c, d) where one cluster has more labelled points. This suggests that balancing of labelled points in practical applications might lead to better accuracy albeit at a high computational cost. The sensitivity to the balancing of labelled points further highlights the importance hierarchical Bayesian methods can tune the τ, α parameters automatically.

5. Conclusions

We have studied the consistency of the probit and one-hot methods for SSL, demonstrating that the combination of ideas from unsupervised learning and supervised learning can lead to consistent labelling of large data sets, given only a few labels. Our theory and numerical results demonstrate that with careful choice of the function of the graph Laplacian appearing in the quadratic penalty, namely the parameters α, τ and ϵ, correct labelling of the data set can be achieved asymptotically.

However, our theory and numerics also indicate that the choice of these parameters is crucial and can lead to failure of the methods. For example, we observed that when ϵ/τ^2 is too large then the methods have a tendency to propagate the majority label rather

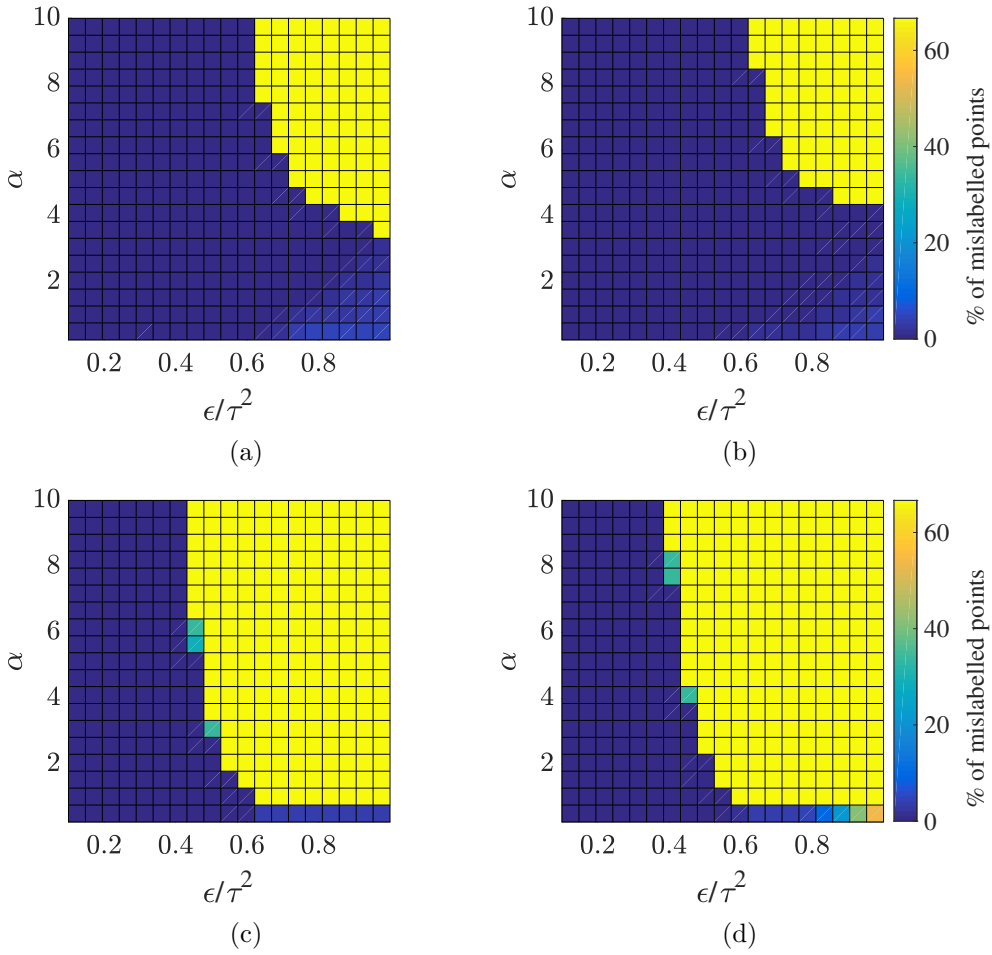
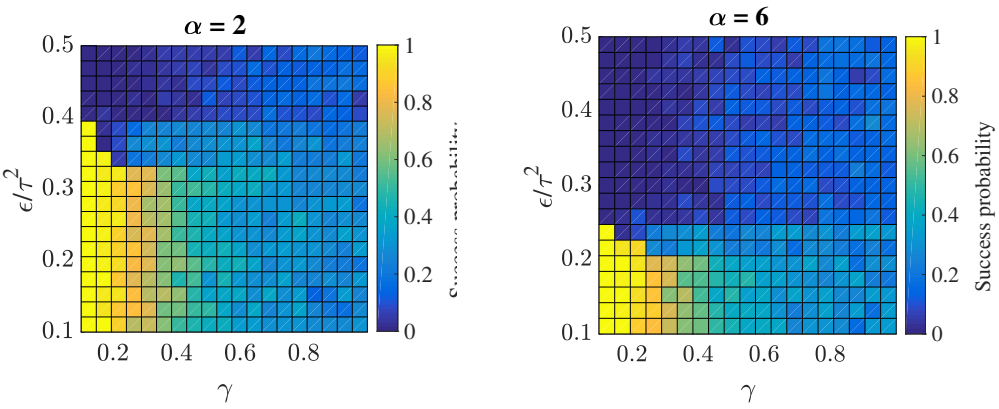


Figure 8: Effect of the number of observed labels on accuracy of one-hot for $n = 10$ and $\gamma = 0.5$ as a function of α and ϵ/τ^2. (a) A single label is observed in each of the three clusters. (b) Three labels are observed in each cluster. (c) Two observations in one cluster and a single observation in the other two. (d) Three observations in one cluster and single observations in the other two.



(a) (b)

Figure 9: Success probability of one-hot with three observations in one cluster and single observations in the other two clusters for $n = 10$, $\alpha = 2$ and 6, and as a function of ϵ/τ^2 and γ. As ϵ/τ^2 grows there appears to be a phase transition where the success probability suddenly drops.

than matching the observed labels. This sensitivity strongly suggests the importance of hierarchical Bayesian techniques which can determine such choices in a data-driven fashion. Proving that hierarchical methods can learn the scaling of τ in terms of ϵ that emerges from our analysis would be of interest.

There are a number of directions in which this work can be taken, including the study of Bayesian posterior consistency, which we undertake in Bertozzi et al. (2020) for the harmonic-function based approach to graph-based SSL, and the study of the limit of large unlabelled data sets in Hoffmann et al. (2020b,a). Another interesting question is a detailed analysis of the majority label propagation phenomenon that is observed when ϵ/τ^2 is too large. Furthermore, throughout this paper we mainly focused on the setting where the ground truth function $u^\dagger$ is consistent with the clustering and assigns the same label to all points within the same cluster. Then the question arises, how do probit and one-hot behave when mislabelled points are present in the data or $u^\dagger$ assigns more than one class to a cluster. It would also be interesting to study different clustering assumptions, such as those arising from the stochastic block models of Lei and Rinaldo (2015).

Acknowledgments

The authors are grateful to Nicolás García-Trillos, Mark Girolami and Omiros Papaspiliopoulos for helpful discussions about the probit methodology and spectral clustering. FH is partially supported by Caltech's von Kármán postdoctoral instructorship. BH is supported in part by an NSERC PDF fellowship. AMS is grateful to AFOSR (grant FA9550-17-1-0185) and NSF (grant DMS 18189770) for financial support.

Appendix A. Spectral Analysis of Covariance Operators

Here we study the spectrum of the covariance operators $C_{\tau,0}$ and $C_{\tau,\epsilon}$ and prove Propositions 13 and 14. Throughout this section we use the notation of Subsection 2.5. We start with some preliminary results regarding the spectrum of $C_{\tau,0}$. Recall that Assumption 1 ensures that G_0 consists of $K < N$ disconnected components $\tilde{G}_k$ and that the graph Laplacian restricted to each $\tilde{G}_k$ has a one-dimensional null-space consisting of constant functions on $\tilde{G}_k$.

Lemma 34 *Suppose $G_0 = \{X, W\}$ is a proximity graph satisfying Assumption 1, and let L_0 be a graph Laplacian operator on G_0 as in (7) with $p = q$. Then L_0 is positive semi-definite for any choice of $p \in \mathbb{R}$ and the matrix $C_{\tau,0}^{-1}$ as in (27) is symmetric and strictly positive definite for any values of $\tau^2 > 0$ and $\alpha > 0$.*

Proof Assume $Z = \{\tilde{Z}_1, \tilde{Z}_2, \cdots, \tilde{Z}_K\}$ where $\tilde{Z}_k$ are the collection of nodes in the k^{th} component $\tilde{G}_k$ of G_0 and let $\tilde{L}_k$ denote the graph Laplacian operator on $\tilde{G}_k$ constructed as in (7) by replacing D_0 and W_0 with $\tilde{D}_k$ and $\tilde{W}_k$ the submatrices corresponding to $\tilde{G}_k$.

Then by (Chung, 1997, Lem. 1.7) the matrices $\tilde{L}_k$ are positive semi-definite with eigenvalues $0 = \tilde{\sigma}_{1,k} < \tilde{\sigma}_{2,k} \leq \tilde{\sigma}_{3,k} \leq \cdots \leq \tilde{\sigma}_{N_k,k}$ where $N_k = |\tilde{Z}_k|$. It follows that $L_0 = \mathrm{diag}(\tilde{L}_1, \tilde{L}_2, \cdots, \tilde{L}_K)$ and so L_0 is also positive semi-definite with eigenvalues $0 = \sigma_{1,0} = \sigma_{2,0} = \cdots = \sigma_{K,0} < \sigma_{K+1,0} \leq \sigma_{K+2,0} \leq \cdots \leq \sigma_{N,0}$.

Now suppose $\alpha = 1$. Then it is straightforward to check that $C_{\tau,0}^{-1}$ has eigenvalues $\lambda_{k,0} = (\tau^{-2}\sigma_{k,0} + 1)$, so that $1 = \lambda_{1,0} = \lambda_{2,0} = \lambda_{K,0} < \lambda_{K+1,0} \leq \lambda_{K+2,0} \leq \cdots \leq \lambda_{N,0}$ and so it is strictly positive definite. The case with $\alpha > 0$ then follows since the eigenvalues are simply $\lambda_{k,0} = (\tau^{-2}\sigma_{k,0} + 1)^{\alpha}$. ∎

Lemma 35 *Suppose $\tau^2, \alpha > 0$ and let $\{\lambda_{k,0}\}_{k=1}^N$ and $\{\boldsymbol{\phi}_{k,0}\}_{k=1}^N$ be the eigenvalues and eigenvectors of $C_{\tau,0}^{-1}$ respectively. Then*

$$\mathbf{c}_{j,0} = \sum_{k=1}^{N} \frac{1}{\lambda_{k,0}} (\boldsymbol{\phi}_{k,0})_j \boldsymbol{\phi}_{k,0}. \tag{63}$$

Proof Since $C_{\tau,0}^{-1}$ is self-adjoint and positive it has an eigendecomposition

$$C_{\tau,0}^{-1} = Q \Lambda Q^{-1},$$

where Λ is a diagonal matrix with diagonal elements $\lambda_{k,0}$ and $Q = [\boldsymbol{\phi}_{1,0}, \cdots, \boldsymbol{\phi}_{N,0}]$ is a unitary matrix with columns $\boldsymbol{\phi}_{k,0}$. Substituting this expression in the identity $\mathbf{c}_{j,0} = C_{\tau,0}\mathbf{e}_j$ and noting that $Q^{-1} = Q^T$ then gives

$$\mathbf{c}_{j,0} = Q\Lambda^{-1}Q^T\mathbf{e}_j = \left[\tfrac{1}{\lambda_{1,0}}\boldsymbol{\phi}_{1,0}, \tfrac{1}{\lambda_{2,0}}\boldsymbol{\phi}_{2,0}, \cdots, \tfrac{1}{\lambda_{N,0}}\boldsymbol{\phi}_{N,0}\right] \begin{bmatrix} (\boldsymbol{\phi}_{1,0})_j \\ (\boldsymbol{\phi}_{2,0})_j \\ \vdots \\ (\boldsymbol{\phi}_{N,0})_j \end{bmatrix},$$

which concludes the proof. ∎

A.1. Proof of Proposition 13: Disconnected Clusters

Proposition 36 (Disconnected Clusters) *Let G_0 satisfy Assumption 1 with $K < N$ components $\tilde{G}_k$. Let $\bar{\boldsymbol{\chi}}_k$ be as in (25). Then for $k = 1, \cdots, K$*

$$\lambda_{k,0} = 1 \quad \textit{and} \quad \boldsymbol{\phi}_{k,0} \in \textit{span}\{\bar{\boldsymbol{\chi}}_j\}_{j=1}^{K},$$

and $\lambda_{K+1,0} > 1$.

Proof The statement involving the eigenvalues $\{\lambda_{k,0}\}_{k=1}^{K+1}$ follows from the proof of Lemma 34 and so we only prove the result regarding the eigenfunctions. Observe that since G_0 consists of K disconnected components then $L_0 = \text{diag}(\tilde{L}_1, \cdots, \tilde{L}_K)$ and each block matrix $\tilde{L}_k$ is itself a graph Laplacian. Writing the inner product $\langle \mathbf{x}, L\mathbf{x} \rangle$ in symmetric form as in (8), it is easy to see that the first eigenvector of any graph Laplacian operator L of the form (7) is $D_0^p \mathbf{1}$ for $p = q$ with corresponding zero eigenvalue. Using this fact for the submatrices $\tilde{L}_k$, we infer that $L_0 \bar{\boldsymbol{\chi}}_k = 0$ for $k = 1, \cdots, K$. We now conclude the proof by noting that $C_{\tau,0}^{-1}$ and L_0 have the same eigenvectors. ■

Proposition 37 *Suppose G_0 satisfies Assumption 1. Then*

$$\mathbf{c}_{j,0} = (\bar{\boldsymbol{\chi}}_k)_j \, \bar{\boldsymbol{\chi}}_k + \mathcal{O}(\tau^{2\alpha}) \qquad \forall j \in \tilde{Z}_k.$$

Proof Follows directly from the expansion (63) and the observation that $\lambda_{1,0} = \cdots = \lambda_{K,0} = 1$ while $\lambda_{K+1,0}^{-1} = \mathcal{O}(\tau^{-2\alpha})$. Finally, note that

$$\sum_{\ell=1}^{K} (\bar{\boldsymbol{\chi}}_\ell)_j \bar{\boldsymbol{\chi}}_\ell = (\bar{\boldsymbol{\chi}}_k)_j \, \bar{\boldsymbol{\chi}}_k \qquad \forall j \in \tilde{Z}_k.$$

■

A.2. Proof of Proposition 14: Weakly Connected Clusters

We now analyze the spectrum of the $C_{\tau,\epsilon}$ operators beginning with auxiliary results regarding the matrices W_ϵ and L_ϵ.

Lemma 38 *Let W_ϵ be as in (28) and suppose Assumptions 1 and 2 are satisfied. Let L_ϵ be as in (30). Then there exists $\epsilon_0 > 0$ so that for all $\epsilon \in (0, \epsilon_0)$ the matrix W_ϵ has non-negative weights and the graph Laplacian operator L_ϵ satisfies an expansion of the form*

$$L_\epsilon = L_0 + \sum_{h=1}^{\infty} \epsilon^h L^{(h)} \tag{64}$$

with $\|L^{(h)}\|_2 \in \ell^\infty$.

Note that in the above, the perturbations $L^{(h)}$ are not necessarily of graph Laplacian form.

Proof First, we prove that for sufficiently small ϵ the W_ϵ matrix has non-negative weights and is therefore a well-defined weight matrix for the graph G_ϵ. By (29), we only need to consider indices i, j for which $w_{ij}^{(0)} > 0$ since $w_{ij}^{(h)}$ may be negative for such indices, i.e., by Assumption 2 these are only the edge weights within clusters since perturbed edge weights between different clusters are constrained to be positive. Since $\|W^{(h)}\|_2$ is uniformly bounded we can use the equivalence of the ℓ_2 and ℓ_∞ norms to infer that all entries of $W^{(h)}$ are also uniformly bounded. Then for $i, j \in \tilde{Z}_k$ we have

$$w_{ij}^{(\epsilon)} = w_{ij}^{(0)} + \sum_{h=1}^{\infty} \epsilon^h w_{ij}^{(h)} \geq w_{ij}^{(0)} - \sup_h |w_{ij}^{(h)}| \left(\frac{\epsilon}{1-\epsilon} \right),$$

which is positive for all $i, j \in \tilde{Z}_k$ for sufficiently small $\epsilon \leq \epsilon_k$. Taking the infimum of ϵ_k we get the uniform constant ϵ_0.

It is straightforward to check that $D_\epsilon = D_0 + \sum_{h=1}^{\infty} \epsilon^h D^{(h)}$ where the $D^{(h)} := \mathrm{diag}(W^{(h)}\mathbf{1})$ are the degree matrices of the $W^{(h)}$. Let $d_i^{(0)}, d_i^{(\epsilon)}$ and $d_i^{(h)}$ denote the diagonal entries of D_0, D_ϵ and $D^{(h)}$ respectively. Then for $i = 1, \cdots, N$,

$$d_i^{(\epsilon)} = d_i^{(0)} \left(1 + \sum_{h=1}^{\infty} \epsilon^h \frac{d_i^{(h)}}{d_i^{(0)}} \right).$$

It follows from Assumption 1(b) that the entries $d_{ii}^{(0)}$ of the degree matrix D_0 are strictly positive (i.e., the components $\tilde{G}_k$ are pathwise connected) and so the ratio inside the inner sum is bounded. Thus, we can further write

$$d_i^{(\epsilon)} = d_i^{(0)} \left(1 + \epsilon \sum_{h=0}^{\infty} \epsilon^h \frac{d_i^{(h+1)}}{d_i^{(0)}} \right) = d_i^{(0)} (1 + \epsilon \hat{d}_i).$$

where the entries $\hat{d}_i$ are well-defined since the sum inside the bracket converges for $\epsilon < 1$. Assuming ϵ is sufficiently small so that $\epsilon \hat{d}_i < 1$ we can use the generalized binomial expansion to write for any $\beta \in \mathbb{R}$

$$\begin{aligned} \left(d_i^{(\epsilon)} \right)^\beta &= \left(d_i^{(0)} \right)^\beta \left(1 + \epsilon \tilde{d}_i \right)^\beta \\ &= \left(d_i^{(0)} \right)^\beta \left(1 + \sum_{h=1}^{\infty} \binom{\beta}{h} \epsilon^h \left(\hat{d}_i \right)^h \right). \end{aligned}$$

Thus, for any $p \in \mathbb{R}$ there exists a diagonal matrix $\hat{D}^{(p)}$ so that

$$D_\epsilon^{-p} = D_0^{-p} + \epsilon \hat{D}^{(p)},$$

and the entries of $\hat{D}^{(p)}$ are uniformly bounded for sufficiently small ϵ. Therefore, we can write

$$\begin{aligned}
L_\epsilon &= D_\epsilon^{-p}(D_\epsilon - W_\epsilon)D_\epsilon^{-p} \\
&= D_\epsilon^{-p}\left(D_0 + \sum_{h=1}^{\infty}\epsilon^h D^{(h)} - W_0 - \sum_{h=1}^{\infty}\epsilon^h W^{(h)}\right)D_\epsilon^{-p} \\
&= D_\epsilon^{-p}\left(D_0 - W_0 + \sum_{h=1}^{\infty}\epsilon^h (D^{(h)} - W^{(h)})\right)D_\epsilon^{-p} \\
&= D_\epsilon^{-p}(D_0 - W_0)D_\epsilon^{-p} + \sum_{h=1}^{\infty}\epsilon^h D_\epsilon^{-p}(D^{(h)} - W^{(h)})D_\epsilon^{-p} \\
&= L_0 + \sum_{h=1}^{\infty}\epsilon^h L^{(h)},
\end{aligned}$$

where the $L^{(h)}$ matrices are obtained by gathering the $\mathcal{O}(\epsilon^h)$ terms. Note that since the entries of $W^{(h)}$ and $\hat{D}^{(p)}$ are uniformly bounded then all entries of the $L^{(h)}$ are bounded uniformly from which it follows that $\|L^{(h)}\|_2 \in \ell^\infty$. ■

We now characterize the low-lying eigenvalues and eigenvectors of $C_{\epsilon,\tau}^{-1}$.

Proposition 39 (The Spectrum of $C_{\tau,\epsilon}^{-1}$) *Suppose Assumptions 1 and 2 are satisfied and let $\{\lambda_{j,\epsilon}, \boldsymbol{\phi}_{j,\epsilon}\}$ denote the orthonormal eigenpairs of $C_{\tau,\epsilon}^{-1}$. Then there exists $\epsilon_0 > 0$ so that*

(i) $\lambda_{1,\epsilon} = 1$ *and* $\boldsymbol{\phi}_{1,\epsilon} = \bar{\boldsymbol{\chi}}$, *as in* (26).

(ii) $\forall \epsilon \in (0, \epsilon_0)$, *there exists constants* $\Xi_1(K, \|L^{(1)}\|_2) > 0$ *and* $\Xi_2(K, \sup_{h\geq 2}\|L^{(h)}\|_2, \epsilon_0) > 0$ *independent of* ϵ *so that*

$$\lambda_{k,\epsilon} \leq \left(1 + \Xi_1\epsilon\tau^{-2} + \Xi_2\epsilon^2\tau^{-2}\right)^\alpha, \qquad \forall k \in \{2, \cdots, K\}. \tag{65}$$

(iii) If there exists a uniform constant $\vartheta > 0$ so that $\lambda_{K+1,\epsilon} - 1 \geq \vartheta$ then there exists a constant $\Xi_3(K, \|L^{(1)}1\|_2, \vartheta) > 0$ independent of $\epsilon \in (0, \epsilon_0)$ so that

$$\left|1 - \sum_{j=1}^{K}\langle \boldsymbol{\phi}_{j,\epsilon}, \bar{\boldsymbol{\chi}}_k\rangle^2\right| \leq \Xi_3\epsilon^2 + \mathcal{O}(\epsilon^3), \qquad \forall j \in \{1, \cdots, K\}, \tag{66}$$

with $\bar{\boldsymbol{\chi}}_k$ as in (25).

Proof (i) Follows from the fact that L_ϵ is a graph Laplacian operator with first eigenvalue $\sigma_{1,\epsilon} = 0$ and first eigenvector $\boldsymbol{\phi}_{1,\epsilon} = D_\epsilon^p\mathbf{1}/\|D_\epsilon^p\mathbf{1}\|$.

(ii) Let $\sigma_{j,\epsilon}$ for $j = 1, \cdots, N$ denote the eigenvalues of L_ϵ. By the min-max principle (see Golub and Van Loan, 1996, Thm. 8.1.2)

$$\sigma_{k,\epsilon} = \min_{\mathbb{U}\in\mathbb{V}_k}\max_{\substack{\mathbf{x}\in\mathbb{U} \\ \|\mathbf{x}\|=1}}\langle \mathbf{x}, L_\epsilon\mathbf{x}\rangle, \tag{67}$$

where $\mathbb{V}_k$ is the set of all k-dimensional subsets in $\mathbb{R}^N$. Now for $k \leq K$ take $\mathbb{U} = \text{span}\{\bar{\boldsymbol{\chi}}_j\}_{j=1}^k$. Since the vectors $\bar{\boldsymbol{\chi}}_j$ are by definition orthonormal then $\mathbb{U}$ is a k-dimensional subspace of

$\mathbb{R}^N$. These vectors are also in the null space of L_0 and so we have for $i, j \in \{1, \dots, K\}$

$$\begin{aligned}\left\langle \bar{\boldsymbol{\chi}}_i, L_\epsilon \bar{\boldsymbol{\chi}}_j \right\rangle &= \sum_{h=1}^{\infty} \epsilon^h \langle \bar{\boldsymbol{\chi}}_i, L^{(h)} \bar{\boldsymbol{\chi}}_j \rangle \le \sum_{h=1}^{\infty} \epsilon^h \|L^{(h)}\|_2 \\ &= \epsilon \|L^{(1)}\|_2 + \epsilon^2 \left(\|L^{(2)}\|_2 + \left(\sup_{h=2,3,\cdots} \|L^{(h)}\|_2 \right) \frac{\epsilon}{1-\epsilon} \right).\end{aligned} \tag{68}$$

We can now generalize this bound to all unit vectors $\mathbf{x} \in \mathbb{U}$ to get

$$\langle \mathbf{x}, L_\epsilon \mathbf{x} \rangle \le \Xi_1 \epsilon + \Xi_2 \epsilon^2$$

where

$$\Xi_1 := K^2 \|L^{(1)}\|_2, \qquad \Xi_2 := \frac{K^2}{(1-\epsilon_0)} \left(\sup_{h=2,3,\cdots} \|L^{(h)}\|_2 \right).$$

From (67) we now infer that

$$\sigma_{k,\epsilon} \le \Xi_1 \epsilon + \Xi_2 \epsilon^2, \qquad \forall k \in \{2, \dots, K\}. \tag{69}$$

Then (65) follows by noting that $\lambda_{k,\epsilon} = \tau^{-2\alpha}(\sigma_{k,\epsilon} + \tau^2)^\alpha$.

(iii) Using the fact that $L_0 \bar{\boldsymbol{\chi}}_k = 0$ for $k = 1, \dots, K$ we can write

$$\begin{aligned}\|L_\epsilon \bar{\boldsymbol{\chi}}_k\|^2 = \langle L_\epsilon \bar{\boldsymbol{\chi}}_k, L_\epsilon \bar{\boldsymbol{\chi}}_k \rangle &= \sum_{h=1}^{\infty} \sum_{\ell=1}^{\infty} \epsilon^h \epsilon^\ell \langle L^{(h)} \bar{\boldsymbol{\chi}}_k, L^{(\ell)} \bar{\boldsymbol{\chi}}_k \rangle \\ &\le \epsilon^2 \|L^{(1)}\|_2^2 + \left(\sup_{\ell=2,3,\cdots} \|L^{(\ell)}\|_2^2 \right) \left(\frac{\epsilon^2}{1-\epsilon} \right)^2.\end{aligned}$$

Now let $\bar{\boldsymbol{\chi}}_k = \sum_{j=1}^N q_{kj} \boldsymbol{\phi}_{j,\epsilon}$ and assume $\sigma_{K+1,\epsilon} \ge \vartheta$. Note that $q_{kj} \le 1$ for all $k \in \{1, ..., K\}$ and $j \in Z$ since $\bar{\boldsymbol{\chi}}_k$ is normalized. Then the above calculation yields

$$\langle L_\epsilon \bar{\boldsymbol{\chi}}_k, L_\epsilon \bar{\boldsymbol{\chi}}_k \rangle = \sum_{j=1}^{N} q_{kj}^2 \sigma_{j,\epsilon}^2 \le \epsilon^2 \|L^{(1)}\|_2^2 + \mathcal{O}(\epsilon^4).$$

From this it follows that

$$\begin{aligned}\vartheta^2 \left(1 - \sum_{j=1}^{K} q_{kj}^2 \right) = \vartheta^2 \sum_{j=K+1}^{N} q_{kj}^2 &\le \sum_{j=K+1}^{N} q_{kj}^2 \sigma_{j,\epsilon}^2 \\ &\le \epsilon^2 \|L^{(1)}\|_2^2 - \sum_{j=1}^{K} q_{kj}^2 \sigma_{j,\epsilon}^2 + \mathcal{O}(\epsilon^4).\end{aligned}$$

Now using (69) we obtain

$$\begin{aligned}\left| 1 - \sum_{j=1}^{K} \langle \boldsymbol{\phi}_{j,\epsilon}, \bar{\boldsymbol{\chi}}_k \rangle^2 \right| = \left| 1 - \sum_{j=1}^{K} q_{kj}^2 \right| &\le \frac{1}{\vartheta^2} \left(\epsilon^2 \|L^{(1)}\|_2^2 + \sum_{j=1}^{K} \sigma_{j,\epsilon}^2 + \mathcal{O}(\epsilon^4)) \right) \\ &\le \Xi_3 \epsilon^2 + \mathcal{O}(\epsilon^3).\end{aligned}$$

The desired result follows since $C_{\epsilon,\tau}^{-1}$ has the same eigenfunctions as L_ϵ. ■

The result in part (iii) of Proposition 39 is central to the rest of our arguments as it states that the eigenvectors $\{\boldsymbol{\phi}_{j,\epsilon}\}_{j=1}^K$ and the functions $\{\bar{\boldsymbol{\chi}}_j\}_{j=1}^K$ have nearly the same span for small ϵ provided that L_ϵ has a uniform spectral gap between $\sigma_{K,\epsilon}$ and $\sigma_{K+1,\epsilon}$. We now show that this condition is satisfied under very general conditions.

Proposition 40 (Existence of Spectral Gaps) *Suppose Assumptions 1 and 2 are satisfied. Then*

$$\sigma_{K+1,\epsilon} \geq \theta - \sum_{h=1}^{\infty} \epsilon^h \|L^{(h)}\|_2 \qquad \text{and} \qquad \lambda_{K+1,\epsilon} \geq \tau^{-2\alpha}\left(\tau^2 + \theta - \sum_{h=1}^{\infty} \epsilon^h \|L^{(h)}\|_2\right)^{\alpha},$$

where $\theta > 0$ is the constant appearing in Assumption 1(b).

Proof By the max-min principle

$$\sigma_{k+1,\epsilon} = \max_{\mathbb{U}\in\mathbb{V}_k} \min_{\substack{\mathbf{x}\perp\mathbb{U} \\ \|\mathbf{x}\|=1}} \langle \mathbf{x}, L_\epsilon \mathbf{x}\rangle. \tag{70}$$

where $\mathbb{V}_k$ denotes the set of k-dimensional subspaces of $\mathbb{R}^N$ and $\mathbf{x}\perp\mathbb{U}$ means the vector $\mathbf{x}$ belongs to the orthogonal complement of $\mathbb{U}$.

Now take $\mathbb{U} = \text{span}\{\bar{\boldsymbol{\chi}}_j\}_{j=1}^K$. Let $\mathbf{x}_k \in \mathbb{R}^{N_k}$ denote the restriction of $\mathbf{x}$ to the subset of indices $\tilde{Z}_k$. Then $\mathbf{x}_k^T(\tilde{D}_k^p \mathbf{1}) = 0$, where we recall $\tilde{D}_k$ is the degree matrix of the k-th cluster $\tilde{G}_k$. Now we can write for all $k \in \{1, ..., K\}$,

$$\begin{aligned}
\langle \mathbf{x}, L_\epsilon \mathbf{x}\rangle &= \langle \mathbf{x}, L_0 \mathbf{x}\rangle + \sum_{h=1}^{\infty} \epsilon^h \langle \mathbf{x}, L^{(h)}\mathbf{x}\rangle \\
&= \sum_{k=1}^{K} \langle \mathbf{x}_k, \tilde{L}_k \mathbf{x}_k\rangle + \sum_{h=1}^{\infty} \epsilon^h \langle \mathbf{x}, L^{(h)}\mathbf{x}\rangle \\
&\geq \theta \sum_{k=1}^{K} \langle \mathbf{x}_k, \mathbf{x}_k\rangle + \sum_{h=1}^{\infty} \epsilon^h \langle \mathbf{x}, L^{(h)}\mathbf{x}\rangle \\
&\geq \theta \|\mathbf{x}\|^2 - \sum_{h=1}^{\infty} \epsilon^h \|L^{(h)}\| \|\mathbf{x}\|^2 \\
&= \left(\theta - \sum_{h=1}^{\infty} \epsilon^h \|L^{(h)}\|_2\right) \|\mathbf{x}\|^2 .
\end{aligned}$$

The lower bound on $\sigma_{K+1,\epsilon}$ now follows from (70) while the lower bound on $\lambda_{K+1,\epsilon}$ follows from the observation that $\lambda_{j,\epsilon} = \tau^{-2\alpha}(\sigma_{j,\epsilon} + \tau^2)^\alpha$ from the definition of $C_{\tau,\epsilon}^{-1}$. ■

Example 1 (Perturbed Kernels) *Consider a proximity graph G where the weight matrix W_0 is given by*

$$w_{ij}^{(0)} = \kappa(x_i - x_j), \qquad i, j \in Z,$$

where $\kappa : \mathbb{R}^N \mapsto \mathbb{R}$ is a positive, uniformly bounded, radially symmetric and non-increasing kernel with full support and $\int_{\mathbb{R}^N} \kappa(x)^2 dx < +\infty$. Now let $\tilde{\kappa} : \mathbb{R}^N \mapsto \mathbb{R}$ be another radially symmetric, positive and uniformly bounded function and define the perturbed weight matrix W_ϵ by

$$w_{ij}^{(\epsilon)} = w_{ij}^{(0)} + \epsilon \tilde{\kappa}(x_i - x_j).$$

Then the resulting perturbed graph Laplacian L_ϵ satisfies the conditions of Propositions 39 and 40. As a concrete example take $\kappa(x) = \exp(-\|x\|)$ and take $\tilde{\kappa}(x) = 1$. For more details on applications using this type of weight matrix, see Zhu (2005). ◇

Example 2 (Adding a Few Edges) *As another example consider a weight matrix W_0 of block diagonal form consisting of two matrices W_+ and W_- corresponding to two disjoint and connected subgraphs, i.e,*

$$W_0 = \mathrm{diag}(W_+, W_-)$$

and any pair of nodes, both in W_+ (resp. W_-) are connected by a sequence of edges with strictly positive weights. Since each subgraph is assumed to be connected then the graph G_0 satisfies Assumption 1. Let Z denote the collection of all nodes in the graph and select a subset $\tilde{Z} \subset Z$ such that $2 \le |\tilde{Z}| \le |Z|$ and $\tilde{Z}$ contains at least one point in each of the two disjoint components. Define the matrix

$$w_{ij}^{(1)} = \begin{cases} 1, & i, j \in \tilde{Z}, i \neq j, \\ 0 & \text{otherwise,} \end{cases}$$

and consider the perturbation

$$W_\epsilon = W_0 + \epsilon W^{(1)}.$$

This perturbation corresponds to weak coupling of two disjoint subgraphs by adding edges between subgraphs. Once again, we can directly verify that the resulting perturbed graph Laplacian L_ϵ satisfies the conditions of Propositions 39 and 40. ◇

At the end of this section we use our results on the closeness of the eigenvectors $\{\boldsymbol{\phi}_{j,\epsilon}\}_{j=1}^K$ and $\{\bar{\boldsymbol{\chi}}_j\}_{j=1}^K$ to identify the geometry of the $\mathbf{c}_{j,\epsilon}$ the columns of $C_{\epsilon,\tau}$.

Proposition 41 *Suppose Assumptions 1 and 2 are satisfied. Then*

(a) If $\epsilon = o(\tau^2)$ there exists a constant $\Xi_4 > 0$ independent of ϵ and τ so that

$$\left\| \mathbf{c}_{\ell,\epsilon} - (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k \right\|^2 \le \Xi_4 \left(\frac{\epsilon^2}{\tau^4} + \tau^{4\alpha} + \epsilon^2 \right), \qquad \forall\, \ell \in \tilde{Z}_k.$$

That is, the columns of $C_{\tau,\epsilon}$ have the same geometry as the weighted set functions $\bar{\boldsymbol{\chi}}_k$ when ϵ, τ and ϵ/τ^2 are small.

(b) If $\epsilon/\tau^2 = \beta$ is constant, there exists a constant $\Xi_5 > 0$ independent of ϵ and τ so that

$$\left\| \mathbf{c}_{\ell,\epsilon} - \left[\left(1 - \breve{\beta}\right) (\bar{\boldsymbol{\chi}})_\ell \bar{\boldsymbol{\chi}} + \breve{\beta} (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k \right] \right\|^2 \le \Xi_5 \left(\epsilon^2 + \tau^{4\alpha} \right), \qquad \forall\, \ell \in \tilde{Z}_k,$$

where $\breve{\beta} = (1 + \Xi_1 \beta)^{-\alpha}$ and Ξ_1 is the constant in (65).

Proof Let $P_\epsilon \in \mathbb{R}^{N\times N}$ denote the projection matrix onto $\operatorname{span}\{\boldsymbol{\phi}_{j,\epsilon}\}_{j=1}^K$ and $P_0 \in \mathbb{R}^{N\times N}$ denote the projection onto $\operatorname{span}\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$. Define the residuals

$$\mathbf{r}_k := (I - P_\epsilon)\bar{\boldsymbol{\chi}}_k = \bar{\boldsymbol{\chi}}_k - \sum_{j=1}^K \langle \boldsymbol{\phi}_{j,\epsilon}, \bar{\boldsymbol{\chi}}_k\rangle \boldsymbol{\phi}_{j,\epsilon}\,,$$
$$\mathbf{s}_j := (I - P_0)\boldsymbol{\phi}_{j,\epsilon} = \boldsymbol{\phi}_{j,\epsilon} - \sum_{k=1}^K \langle \boldsymbol{\phi}_{j,\epsilon}, \bar{\boldsymbol{\chi}}_k\rangle \bar{\boldsymbol{\chi}}_k\,.$$

By Proposition 39(iii) there exists a uniform constant $\Xi_6 > 0$ so that

$$\sum_{k=1}^K \|\mathbf{r}_k\|^2 = \sum_{j=1}^K \|\mathbf{s}_j\|^2 = K - \sum_{k=1}^K \sum_{j=1}^K \langle \boldsymbol{\phi}_{j,\epsilon}, \bar{\boldsymbol{\chi}}_k\rangle^2 \le K\Xi_6\epsilon^2\,. \tag{71}$$

Writing $\boldsymbol{\phi}_{j,\epsilon} = P_0\boldsymbol{\phi}_{j,\epsilon} + \mathbf{s}_j$ for any $j \in \{1,\dots,N\}$ and using the fact that the $\{\boldsymbol{\phi}_{j,\epsilon}\}_{j=1}^K$ and $\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$ are unit vectors, we can estimate

$$\begin{aligned}
\left\|\sum_{j=1}^K (\boldsymbol{\phi}_{j,\epsilon})_\ell \boldsymbol{\phi}_{j,\epsilon} - \sum_{k=1}^K (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k\right\|^2 &= \left\|\sum_{j=1}^K (\boldsymbol{\phi}_{j,\epsilon})_\ell P_0\boldsymbol{\phi}_{j,\epsilon} - \sum_{k=1}^K (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k + \sum_{j=1}^K (\boldsymbol{\phi}_{j,\epsilon})_\ell \mathbf{s}_j\right\|^2 \\
&\le 2\left\|\sum_{j=1}^K (\boldsymbol{\phi}_{j,\epsilon})_\ell P_0\boldsymbol{\phi}_{j,\epsilon} - \sum_{k=1}^K (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k\right\|^2 + 2K\sum_{j=1}^K \|\mathbf{s}_j\|^2 \\
&= 2\left\|\sum_{k=1}^K \left[\left(\sum_{j=1}^K (\boldsymbol{\phi}_{j,\epsilon})_\ell \langle \bar{\boldsymbol{\chi}}_k, \boldsymbol{\phi}_{j,\epsilon}\rangle\right) - (\bar{\boldsymbol{\chi}}_k)_\ell\right]\bar{\boldsymbol{\chi}}_k\right\|^2 + 2K\sum_{j=1}^K \|\mathbf{s}_j\|^2 \\
&= 2\left\|\sum_{k=1}^K ((P_\epsilon - I)\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k\right\|^2 + 2K\sum_{j=1}^K \|\mathbf{s}_j\|^2 \\
&\le 2K\sum_{k=1}^K \|(P_\epsilon - I)\bar{\boldsymbol{\chi}}_k\|^2 + 2K\sum_{j=1}^K \|\mathbf{s}_j\|^2 \\
&= 2K\sum_{k=1}^K \|\mathbf{r}_k\|^2 + 2K\sum_{j=1}^K \|\mathbf{s}_j\|^2 \le 4K^2\Xi_6\epsilon^2\,,
\end{aligned} \tag{72}$$

where the last inequality follows from (71).

Now recall the expansion (21) for columns of $C_{\epsilon,\tau}$,

$$\mathbf{c}_{\ell,\epsilon} = \sum_{j=1}^N \frac{1}{\lambda_{j,\epsilon}} (\boldsymbol{\phi}_{j,\epsilon})_\ell \boldsymbol{\phi}_{j,\epsilon}.$$

For (a), we want to estimate

$$\begin{aligned}
\left\|\mathbf{c}_{\ell,\epsilon} - \sum_{k=1}^K (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k\right\|^2 \le 3\left\|\sum_{j=K+1}^N \frac{1}{\lambda_{j,\epsilon}} (\boldsymbol{\phi}_{j,\epsilon})_\ell \boldsymbol{\phi}_{j,\epsilon}\right\|^2 + 3\left\|\sum_{j=1}^K \left(\frac{1}{\lambda_{j,\epsilon}} - 1\right)(\boldsymbol{\phi}_{j,\epsilon})_\ell \boldsymbol{\phi}_{j,\epsilon}\right\|^2 \\
+ 3\left\|\sum_{j=1}^K (\boldsymbol{\phi}_{j,\epsilon})_\ell \boldsymbol{\phi}_{j,\epsilon} - \sum_{k=1}^K (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k\right\|^2
\end{aligned} \tag{73}$$

The last term is of order ϵ^2 by (72). The first term can be controlled using Proposition 40 together with the same bound as in (68),

$$\begin{aligned}\left\| \sum_{j=K+1}^{N} \frac{1}{\lambda_{j,\epsilon}} (\boldsymbol{\phi}_{j,\epsilon})_\ell \boldsymbol{\phi}_{j,\epsilon} \right\|^2 &\le \left(\frac{(N-K)}{\lambda_{K+1,\epsilon}} \right)^2 \\ &\le (N-K)^2 \tau^{4\alpha} \left(\tau^2 + \theta - \epsilon \|L^{(1)}\|_2 - \epsilon^2 \left(\sup_{h \ge 2} \|L^{(h)}\|_2 \right) \right)^{-2\alpha} \\ &\le (N-K)^2 \theta^{-2\alpha} \tau^{4\alpha} + \mathcal{O}\left(\tau^{4\alpha} \left(\tau^2 + \epsilon \right) \right),\end{aligned}$$

where $\theta > 0$ is the constant appearing in Assumption 1(b). Next, when ϵ/τ^2 is small, we can estimate the second term in the right-hand side of (73) using Proposition 39(ii): recall that $\lambda_{1,\epsilon} = 1$, and so

$$\left| \frac{1}{\lambda_{j,\epsilon}} - 1 \right| = \frac{|1 - \lambda_{j,\epsilon}|}{|\lambda_{j,\epsilon}|} \le |1 - \lambda_{j,\epsilon}| \le \alpha \Xi_1 \frac{\epsilon}{\tau^2} + \mathcal{O}\left(\epsilon^2 \tau^{-4} \right) \qquad \text{for } j \in \{2, ..., K\}, \tag{74}$$

and therefore

$$\begin{aligned}\left\| \sum_{j=1}^{K} \left(\frac{1}{\lambda_{j,\epsilon}} - 1 \right) (\boldsymbol{\phi}_{j,\epsilon})_\ell \boldsymbol{\phi}_{j,\epsilon} \right\|^2 &\le K \sum_{j=1}^{K} \left| \frac{1}{\lambda_{j,\epsilon}} - 1 \right|^2 \left| (\boldsymbol{\phi}_{j,\epsilon})_\ell \right|^2 \|\boldsymbol{\phi}_{j,\epsilon}\|^2 \\ &\le \alpha^2 K^2 \Xi_1^2 \left(\frac{\epsilon}{\tau^2} \right)^2 + \mathcal{O}\left(\epsilon^3 \tau^{-6} \right).\end{aligned}$$

Putting the above estimates together, we can find a constant $\Xi_7 = \Xi_7(N, K, \theta, \alpha, \epsilon_0) > 0$ independent of ϵ and τ such that for all $\ell \in Z$,

$$\left\| \mathbf{c}_{\ell,\epsilon} - \sum_{k=1}^{K} (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k \right\|^2 \le \Xi_7 \left(\tau^{4\alpha} + \left(\frac{\epsilon}{\tau^2} \right)^2 + \epsilon^2 \right) + \mathcal{O}\left(\tau^{4\alpha} \left(\tau^2 + \epsilon \right) \right) \Big) + \mathcal{O}\left(\left(\frac{\epsilon}{\tau^2} \right)^3 \right).$$

Finally, note that for each $\ell \in Z$, there exists a unique $k_0 \in \{1, ..., K\}$ such that $\ell \in \tilde{Z}_{k_0}$. Then

$$\sum_{k=1}^{K} (\bar{\boldsymbol{\chi}}_k)_\ell \bar{\boldsymbol{\chi}}_k = \left(\bar{\boldsymbol{\chi}}_{k_0} \right)_\ell \bar{\boldsymbol{\chi}}_{k_0},$$

which concludes the proof of statement (a).

To prove (b), the argument is similar. For $\beta := \epsilon/\tau^2$, we have

$$\breve{\beta} = (1 + \beta \Xi_1)^{-\alpha} > 0.$$

Then, instead of (74), we use the bound in Proposition 39(ii) to estimate

$$\begin{aligned}\left| \frac{1}{\lambda_{j\epsilon}} - \breve{\beta} \right| = \frac{|1 - \breve{\beta} \lambda_{j,\epsilon}|}{|\lambda_{j,\epsilon}|} &\le |1 - \breve{\beta} \lambda_{j,\epsilon}| = \breve{\beta} \left| \lambda_{j,\epsilon} - \frac{1}{\breve{\beta}} \right| \\ &\le \breve{\beta} \left| \left(\breve{\beta}^{-1/\alpha} + \epsilon \beta \Xi_2 \right)^{\alpha} - \breve{\beta}^{-1} \right| = \frac{\alpha \beta \Xi_2}{(1 + \beta \Xi_1)} \epsilon + \mathcal{O}(\epsilon^2) \qquad \text{for } j \in \{2, ..., K\}.\end{aligned}$$

Therefore, our goal is to control

$$\begin{aligned}
&\left\|\mathbf{c}_{\ell,\epsilon} - (1-\breve{\beta})\,(\boldsymbol{\phi}_{1,\epsilon})_\ell\,\boldsymbol{\phi}_{1,\epsilon} - \breve{\beta}\sum_{k=1}^{K}(\bar{\boldsymbol{\chi}}_k)_\ell\,\bar{\boldsymbol{\chi}}_k\right\|^2 \\
&\qquad\le 3\left\|\sum_{j=K+1}^{N}\frac{1}{\lambda_{j,\epsilon}}\,(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon}\right\|^2 \\
&\qquad\quad + 3\left\|\sum_{j=1}^{K}\frac{1}{\lambda_{j,\epsilon}}\,(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon} - (1-\breve{\beta})\,(\boldsymbol{\phi}_{1,\epsilon})_\ell\,\boldsymbol{\phi}_{1,\epsilon} - \breve{\beta}\left(\sum_{j=1}^{K}(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon}\right)\right\|^2 \\
&\qquad\quad + 3\left\|\breve{\beta}\left(\sum_{j=1}^{K}(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon} - \sum_{k=1}^{K}(\bar{\boldsymbol{\chi}}_k)_\ell\,\bar{\boldsymbol{\chi}}_k\right)\right\|^2 \qquad (75)\\
&\qquad= 3\left\|\sum_{j=K+1}^{N}\frac{1}{\lambda_{j,\epsilon}}\,(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon}\right\|^2 \\
&\qquad\quad + 3\left\|\sum_{j=2}^{K}\left(\frac{1}{\lambda_{j,\epsilon}} - \breve{\beta}\right)(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon}\right\|^2 \\
&\qquad\quad + 3\breve{\beta}^2\left\|\sum_{j=1}^{K}(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon} - \sum_{k=1}^{K}(\bar{\boldsymbol{\chi}}_k)_\ell\,\bar{\boldsymbol{\chi}}_k\right\|^2 .
\end{aligned}$$

Thanks to the previous estimate, the second term can be bounded by

$$\left\|\sum_{j=2}^{K}\left(\frac{1}{\lambda_{j,\epsilon}} - \breve{\beta}\right)(\boldsymbol{\phi}_{j,\epsilon})_\ell\,\boldsymbol{\phi}_{j,\epsilon}\right\|^2 \le (K-1)^2\left(\frac{\alpha\beta\Xi_2}{(1+\beta\Xi_1)}\right)^2\epsilon^2 + \mathcal{O}(\epsilon^3)\,.$$

The first and last terms in (75) can be estimated as for part (a), and so we can find a constant $\Xi_8 = \Xi_8(N,K,\theta,\alpha,\epsilon_0) > 0$ independent of ϵ and τ such that for all $\ell \in Z$,

$$\begin{aligned}
&\left\|\mathbf{c}_{\ell,\epsilon} - (1-\breve{\beta})\,(\boldsymbol{\phi}_{1,\epsilon})_\ell\,\boldsymbol{\phi}_{1,\epsilon} - \breve{\beta}\sum_{k=1}^{K}(\bar{\boldsymbol{\chi}}_k)_\ell\,\bar{\boldsymbol{\chi}}_k\right\|^2 \\
&\qquad\le \Xi_8\left(\tau^{4\alpha}+\epsilon^2\right) + \mathcal{O}\left(\tau^{4\alpha}\left(\tau^2+\epsilon\right)\right) + \mathcal{O}\left(\epsilon^3\right).
\end{aligned}$$

From Proposition 39(i), we know that $\boldsymbol{\phi}_{1,\epsilon} = D_\epsilon^p\mathbf{1}/\|D_\epsilon^p\mathbf{1}\|$. Further, using the expansion $D_\epsilon^p = D_0^p + \epsilon\hat{D}^{(p)}$ derived in the proof of Lemma 38, we can simplify the middle term in the expression on the left-hand side above,

$$(1-\breve{\beta})\,(\boldsymbol{\phi}_{1,\epsilon})_\ell\,\boldsymbol{\phi}_{1,\epsilon} = (1-\breve{\beta})(\bar{\boldsymbol{\chi}})_\ell\bar{\boldsymbol{\chi}} + \mathcal{O}(\epsilon)\,.$$

This concludes the proof of Proposition 41. ■

References

Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26(2):445–469, 2005.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 585–591, 2001.

Mikhail Belkin and Partha Niyogi. Convergence of laplacian eigenmaps. In *Proceedings of the 19th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 129–136, 2007.

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

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

Andrea L Bertozzi and Arjuna Flenner. Diffuse interface models on graphs for classification of high dimensional data. *Multiscale Modeling & Simulation*, 10(3):1090–1118, 2012.

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

Andrea L Bertozzi, Bamdad Hosseini, Hao Li, Kevin Miller, and Andrew M Stuart. Posterior consistency of semi-supervised regression on graphs. arXiv preprint:2007.12809, 2020.

Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning*, pages 19–26, 2001.

Vladimir I Bogachev. *Differentiable measures and the Malliavin calculus*, volume 164 of *Mathematical Sureverys and Monographs*. AMS, Providence, 2010.

Emmanuel J Candès and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics*, 48 (1):27–42, 2020.

Victor Chen, Matthew M Dunlop, Omiros Papaspiliopoulos, and Andrew M Stuart. Dimension-robust mcmc in bayesian inverse problems. arXiv preprint:1803.03344, 2018.

Fan RK Chung. *Spectral graph theory*. Number 92 in CBMS Regional Conference Series in Mathematics. AMS, Providence, 1997.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, Cambridge, third edition, 2009.

Matthew M Dunlop, Dejan Slepčev, Andrew M Stuart, and Matthew Thorpe. Large data and zero noise limits of graph-based semi-supervised learning algorithms. *Applied and Computational Harmonic Analysis*, 49(2):655–697, 2020.

Nicolás García-Trillos and Dejan Slepčev. A variational approach to the consistency of spectral clustering. *Applied and Computational Harmonic Analysis*, 45(2):239–281, 2018.

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

Gene H Golub and Charles F Van Loan. *Matrix computations*. The Johns Hopkins University Press, Baltimore, third edition, 1996.

Franca Hoffmann, Bamdad Hosseini, Assad A Oberai, and Andrew M Stuart. Spectral analysis of weighted Laplacians arising in data clustering. arXiv preprint:1909.06389, 2020a.

Franca Hoffmann, Bamdad Hosseini, Assad A Oberai, and Andrew M Stuart. Consistency of graphical semi-supervised learning algorithms in the continuum limit: The probit method. In preparation, 2020b.

Georgios Kostopoulos, Stamatis Karlos, Sotiris Kotsiantis, and Omiros Ragos. Semi-supervised regression: A recent review. *Journal of Intelligent & Fuzzy Systems*, 35(2): 1483–1500, 2018.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

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

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, pages 849–856, 2001.

Andràs Prékopa. Logarithmic concave measures and related topics. In *Stochastic Programming*, pages 63–82, 1980.

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

Carl E Rasmussen and Christopher KI Williams. *Gaussian Processes for Machine Learning*. MIT press, Cambridge, 2006.

Adrien Saumard and Jon A Wellner. Log-concavity and strong log-concavity: a review. *Statistics Surveys*, 8:45, 2014.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, Cambridge, 2002.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Dejan Slepčev and Matthew Thorpe. Analysis of p-Laplacian regularization in semisupervised learning. *SIAM Journal on Mathematical Analysis*, 51(3):2085–2120, 2019.

Alex J Smola and Bernhard Schölkopf. On a kernel-based method for pattern recognition, regression, approximation, and operator inversion. *Algorithmica*, 22(1-2):211–231, 1998.

Daniel A Spielman and Shang-Hua Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *Proceedings of the 37th Conference on Foundations of Computer Science*, pages 96–105, 1996.

Daniel A Spielman and Shang-Hua Teng. Spectral partitioning works: Planar graphs and finite element meshes. *Linear Algebra and its Applications*, 421(2-3):284–305, 2007.

Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, 2005.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116 (29):14516–14525, 2019.

Ambuj Tewari and Peter L Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

Vladimir N Vapnik. *Statistical learning theory*. Springer, New York, 1995.

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

Ulrike Von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.

Mingrui Wu and Bernhard Schölkopf. Transductive classification via local learning regularization. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pages 628–635, 2007.

Qiang Wu and Ding-Xuan Zhou. Analysis of support vector machine classification. *Journal of Computational Analysis & Applications*, 8(2), 2006.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.

Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning*, pages 912–919, 2003a.

Xiaojin Zhu, John Lafferty, and Zoubin Ghahramani. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In *ICML workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, 2003b.

Xiaojin Jerry Zhu. Semi-supervised learning literature survey. Technical Report TR1530, University of Wisconsin-Madison, Computer Sciences Department, 2005.