POSTERIOR CONSISTENCY OF SEMI-SUPERVISED REGRESSION ON GRAPHS [*]

ANDREA L. BERTOZZI[‡], BAMDAD HOSSEINI[†], HAO LI[‡], KEVIN MILLER[‡], AND ANDREW M. STUART[†]

Abstract. Graph-based semi-supervised regression (SSR) is the problem of estimating the value of a function on a weighted graph from its values (labels) on a small subset of the vertices. This paper is concerned with the consistency of SSR in the context of classification, in the setting where the labels have small noise and the underlying graph weighting is consistent with well-clustered nodes. We present a Bayesian formulation of SSR in which the weighted graph defines a Gaussian prior, using a graph Laplacian, and the labeled data defines a likelihood. We analyze the rate of contraction of the posterior measure around the ground truth in terms of parameters that quantify the small label error and inherent clustering in the graph. We obtain bounds on the rates of contraction and illustrate their sharpness through numerical experiments. The analysis also gives insight into the choice of hyperparameters that enter the definition of the prior.

Key words. Semi-supervised learning, classification, consistency, graph Laplacian, Bayesian inference.

AMS subject classifications. 62H30, 62F15, 68R10, 68T10, 68Q87.

1. Introduction.

Semi-supervised learning (SSL) is the problem of labeling all points within a dataset (the *unlabeled data*) by combining knowledge of a subset of noisy observed labels (the *labeled data*); this is done by exploiting correlations and geometric information present in the dataset combined with label information. We study this problem in the framework of Bayesian inverse problems (BIPs), building on a widely adopted semi-supervised regression (SSR) approach to SSL developed in the machine learning community. In this context the Bayesian formulation has a novel structure in which the unlabeled data defines the prior distribution and the labeled data defines the likelihood. The goal of this article is to study posterior consistency; that is, the contraction of the resulting Bayesian posterior distribution onto the ground truth solution in certain parametric limits related to parameters underlying our model. We adopt ideas from spectral clustering in unsupervised learning to construct and analyze the prior arising from a similarity graph constructed from the unlabeled data. This prior information interacts with the labeled data via the likelihood. When the prior information (from the unlabeled data) and the likelihood (from the labeled data) complement each other, then a form of Bayesian posterior consistency can be achieved and the posterior measure on the predicted labels contracts around the ground truth. Furthermore our analysis elucidates how hyperparameter choices in the prior, quantitative measures of clustering in the dataset and the noise in labels combine to affect the contraction rates of the posterior. In the following three subsections we review relevant literature, formulate the problem mathematically and describe our contributions.

[*] Submitted to the editors DATE.

Funding: This work is supported by NSF grant DMS 1818977, AFOSR grant FA9550-17-1-0185, NSERC PDF fellowship, a Caltech Von Kármán instructorship, DOD NDSEG Fellowship, and DARPA grant FA8750-18-2-0066.

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

[‡] Department of Mathematics, University of California, Los Angeles, Los Angeles, CA (bertozzi@math.ucla.edu, lihao0809@math.ucla.edu, millerk22@math.ucla.edu).

1.1. Relevant Literature. Many approaches to SSL and SSR 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 [46] and [24] for, respectively, the state-of-the-art in 2005 and a more recent appraisal of the field.

The consistency of supervised learning and regression is well-developed; see [34] for a literature review, as well as the preceding work in [32,33,41] which establish the problem in the framework of Vapnik [37]. All of this work on supervised classification focuses on the large data/large number of features setting, and often considers only linearly separable unlabeled data. Therefore, these previous works do not leverage the power of graph-based techniques to extract geometric information in large unlabeled datasets, a primary feature of the SSR problems studied in this work.

Graph-based techniques are widely used in unsupervised learning [3,38], a subject that has seen significant analysis in relation to consistency. The papers [30,31] perform a careful analysis of the spectral gaps of graph Laplacians resulting from clustered data, studying recursive methods for multi-class clustering. The paper [28] introduced an approach for the analysis of multi-class unsupervised learning based on perturbations of a perfectly clustered case. The paper [39] introduced the idea of studying the consistency of spectral clustering in the limit of large i.i.d. datasets in which the graph Laplacians converge to a limiting integral operator. The articles [14, 15] took this idea further by proving the convergence of graph Laplacian operators to local differential operators by controlling the local connectivity of the graph as a function of the number of vertices.

In this paper our focus is on transductive SSL [24] in the framework of the influential papers [44,45] where the categorical labels $\{1, \dots, M\}$ are embedded in $\mathbb{R}^M$ and the SSR approach to SSL is adopted. Bertozzi and Flenner [5] introduced an interesting relaxation of this assumption, by means of a Ginzburg-Landau penalty term which favors real-values close to ± 1 but does not enforce the categorical values ± 1 exactly. In contrast to these relaxations, the probit approach to classification, described in the classic text on Gaussian process regression [29] and analyzed in [21] in the context of SSL, works directly with the categorical labels and does not rely on the embedding step.

The idea of regularization by graph Laplacians for SSL was developed in different contexts such as manifold regularization [4], Tikhonov regularization [2] and local learning regularization [40] as well as more recent articles focusing on large data settings [18, 19]. However, while graph regularization methods are widely applied in practice the rigorous analysis of their properties, and in particular asymptotic consistency and posterior contraction rates, are not well-developed within the context of SSL and SSR. Indeed, to the best of our knowledge the Bayesian consistency of SSR has not been analyzed. Studying SSL/SSR in a Bayesian setting introduces new challenges that require careful consideration about assumptions regarding graph structure and statistical properties of the resulting model [23]. We build on the spectral analysis of the graph Laplacian introduced in [28] to study unsupervised learning, and refined in [21] to study the consistency of optimization-based approaches to binary and one-hot SSL.

The subject of Bayesian posterior consistency is aimed at reconciling the large data limits of frequentist and Bayesian approaches to statistical inference problems. Early influential works in this field concentrated on negative results concerning the Bayesian nonparametric setting where the prior and likelihood were inconsistent [11]. Subsequent work in this area concentrated on positive results, demonstrating that minimax rates of convergence can be obtained within the Bayesian setting [16, 35] by studying posterior measure concentration through Bernstein-Von Mises-type theorems [13,35] provided that priors are constructed carefully. The celebrated paper [7] demonstrates how large data and small noise limits are intimately related, and this link underpins subsequent studies of inverse problems from the perspective of Bayesian posterior consistency. This line of work was initiated in the paper [36] where the small noise limit of linear inverse problems was studied. A number of papers in this area followed [1,27] and it is currently an active research area, particularly in relation to nonlinear inverse problems [17].

In some problems optimization approaches rather than fully Bayesian approaches are adopted, and the study of consistency for inverse problems in this setting is overviewed in [12]. Linking this to maximum a posteriori (MAP) estimators for inverse problems was a subject developed in [9] and the study of consistency for MAP estimators in semi-supervised learning, and in particular use of the probit likelihood model, is undertaken in [21].

1.2. Problem Setup. Consider a set of nodes $Z = \{1, \cdots, N\}$ and an associated set of *feature vectors* $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$. Each feature vector $\mathbf{x}_j$ 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 a matrix in $\mathbb{R}^{d\times N}$ with columns given by $\mathbf{x}_j$. We refer to X as the *unlabeled data*. Throughout this article we assume that every element of Z belongs to one of M classes and employ the one-hot encoding to represent the label of each point. More precisely, we assume there exists a function $l : Z \mapsto \{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$ where the $\mathbf{e}_j \in \mathbb{R}^M$ are the standard coordinate vectors. A point $j \in Z$ then belongs to class m if $l(j) = \mathbf{e}_m$.

Now let $Z' \subseteq Z$ be a subset of $J \leq N$ nodes and define a function $Y : Z' \mapsto \mathbb{R}^M$, noting that this may also be viewed as a matrix $Y \in \mathbb{R}^{M\times J}$. The columns of Y are denoted by $\{\mathbf{y}'_1, \cdots, \mathbf{y}'_J\}$ and comprise a collection of *noisy observed labels* on Z'; in practice, we use $\mathbf{y}'_j \in \{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$, the one-hot vectors, or small noisy perturbations of this setting. We refer to Y as the *labeled data*. Underlying this paper is the assumption that the labeled data is determined by a generative model of the form

$$Y = U^\dagger H^T + \gamma\eta, \tag{1.1}$$

here $U^\dagger \in \mathbb{R}^{M\times N}$ is the *ground truth latent variable* that gives the true labels of all of the vertices in Z, $H \in \mathbb{R}^{J\times N}$ is the matrix obtained by removing the $Z \setminus Z'$ rows of the identity matrix $I_N \in \mathbb{R}^{N\times N}$ and $\eta \in \mathbb{R}^{M\times J}$ is a matrix with independent standard Gaussian entries, i.e., $\eta_{mj} \overset{iid}{\sim} \mathcal{N}(0,1)$. The parameter $\gamma > 0$ is the standard deviation of the observation noise. It is instructive to think of the columns of $U^\dagger$ as being chosen from $\{\mathbf{e}_1, \cdots, \mathbf{e}_M\}$, although generalizations of this setting are possible.

The model (1.1) casts the SSL problem of inferring the true labels on Z as the SSR problem of finding $U^\dagger$, adopting the terminology of [24]: our modeling assumption makes the observations Y real-valued, rather than categorical as in classification, and therefore is considered a regression problem. The SSR problem is very ill-posed, requiring the learning of NM parameters from JM noisy data points, since we typically have far fewer labels than the total number of unlabeled data points, i.e. $J \ll N$. The labeled data may be viewed as providing prior information that renders this ill-posed problem tractable. To this end we formulate SSR in the framework of Bayesian inversion [8,10,22].

The main goal of this article is to analyze the consistency of the Bayesian SSR problem by identifying the conditions under which the posterior measure μ^Y (defined in (2.7) below) contracts around the ground truth matrix $U^\dagger$ in (1.1). Formally, we define the following functional as a measure of posterior contraction

$$\mathcal{I} := \mathbb{E}_{Y|U^\dagger}\mathbb{E}_{\mu^Y}\left\|U - U^\dagger\right\|_F^2, \tag{1.2}$$

where the inner expectation is with respect to the posterior measure μ^Y on U while the outer expectation is with respect to the law of $Y|U^\dagger$ following (1.1); $\|\cdot\|_F$ denotes the Frobenius norm. With this notation, our aim is to solve the following problem:

Problem 1.1 (Posterior consistency of Bayesian SSR). Under what conditions on the graph G, the labeled set Z', the ground truth $U^\dagger$ and the hyperparameters τ, α entering the definition of the prior can we ensure that $\mathcal{I} \downarrow 0$ as the noise-level γ in the unlabeled data, and some measure ϵ of closeness to perfect clustering in the labeled data, tend to zero.

Indeed we will find explicit bounds on $\mathcal{I}$ which give consistency in the limit $(\epsilon, \gamma) \to 0$ and reveal the role of parameter choices for τ, α in the form of the contraction rate. Our bounds are applicable for small values of γ, τ, ϵ and not just in the asymptotic regimes where $(\gamma, \tau, \epsilon) \to 0$.

1.3. Main Contributions.

We study posterior contraction, as measured by the quantity $\mathcal{I}$. In the theory we develop, the quantity of labeled data and unlabeled data will be fixed, a practically useful setting in which to study algorithms based around SSR. The prior we use is a discrete analog of the Matérn prior with graph Laplacian used in place of the continuum Laplacian in the differential operator formulation popularized in [26]. In this interpretation τ is an inverse length-scale and α controls the regularity of the prior; details are given in the next section. The parameter γ is the noise standard deviation in (1.1) and the parameter ϵ is defined formally through the notion of a weakly connected graph as introduced in [28] and used in [21]:

Definition 1.2 (Weakly connected graph). *Let $0 < \epsilon \ll 1$, then a graph $G = \{Z, W\}$ is weakly connected with K clusters if it consists of pathwise connected components $\widetilde{G}_k = \{\widetilde{Z}_k, \widetilde{W}_k\}$ for $k = 1, \dots, K$ so that the edge weights between elements in different $\widetilde{G}_k$ are $\mathcal{O}(\epsilon)$. In other words, up to a reordering of Z, the matrix W is an $\mathcal{O}(\epsilon)$ perturbation of a block diagonal weight matrix, and the graph Laplacian associated with each block has a one-dimensional null-space.*

The following informal theorem is stated with precise conditions as Corollary 3.16 which itself follows from Theorem 3.12, both stated and proved in Section 3.

Main Theorem. *Let $G = \{Z, W\}$ be weakly connected with K components $\widetilde{G}_k$ and perturbation parameter $0 < \epsilon \ll 1$ as in Definition 1.2. Suppose that the rows of the ground truth matrix $U^\dagger \in \mathbb{R}^{M \times N}$ belong to the span of the indicator functions of the $\widetilde{G}_k$ and fix $\alpha > 0$ and fix τ so that*

$$\epsilon = \epsilon_0 \tau^{\max\{2, 2\alpha\}}.$$

Then, for appropriately chosen ϵ_0, there exists $\Xi > 0$, independent of ϵ and γ, so that

$$\mathcal{I} \leq \Xi \max\left\{\gamma^2, \epsilon^{\min\{1,\alpha\}}\right\}.$$

Let us give insight into this theorem. The parameters ϵ and γ are inherent to the specific SSR problem and the dataset at hand. Broadly speaking ϵ is a geometric property of the point cloud X of unlabeled data and γ is the noise standard deviation

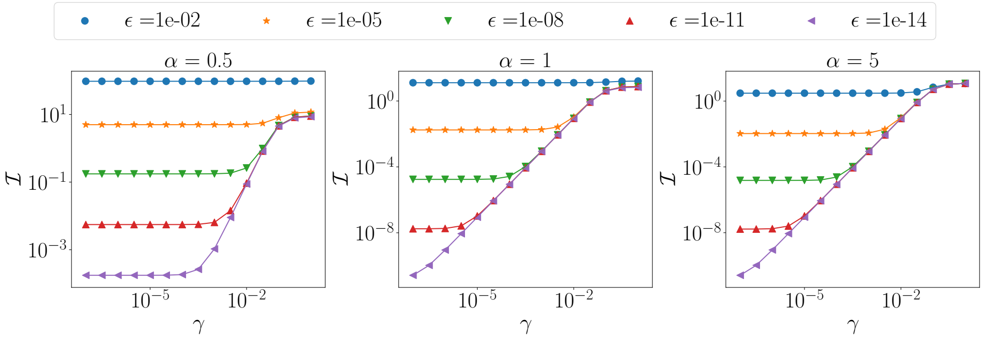


FIG. 1. *A numerical demonstration of the Main Theorem on a synthetic dataset (detailed in Subsection 4.1). Details of this experiment are described in Section 4. The value of $\mathcal{I}$ reduces with γ up to the point where $\gamma^2 \approx \epsilon^{\min\{1,\alpha\}}$ where the errors saturate as predicted by the upper bound in the Main Theorem. Smaller values of ϵ result in smaller values of $\mathcal{I}$ that indicates higher concentration of posterior probability mass around the ground truth $U^\dagger$.*

of the labels. Hence these parameters are fixed, although they are generally unknown. Then the Main Theorem implies the following:

- If $\epsilon^{\min\{1,\alpha\}} \leq \gamma^2$, then the measurement noise dominates over the measure of closeness to perfect clustering and posterior contraction is controlled by the γ parameter.
- If $\gamma^2 < \epsilon^{\min\{1,\alpha\}}$, then the measure of closeness to perfect clustering is dominant in comparison to the measurement noise, and posterior contraction is controlled by the ϵ parameter.
- In the latter case we also observe that choosing $\alpha < 1$ gives a sublinear contraction rate in ϵ while a linear rate is achieved if $\alpha \geq 1$. Thus it is preferable to tune (α, τ^2) so that $\alpha \geq 1$ and $\tau^2 = \mathcal{O}(\epsilon^{1/\alpha})$. For reasons related to the large data limit $N \to \infty$ it is natural to choose $\alpha > \frac{d}{2}$ and since d is typically larger than 2, this enforces $\alpha > 1$; see [20].

These insights are also supported by our numerical experiments in Section 4; furthermore these experiments also verify the sharpness of the upper bound in the Main Theorem. As a prelude to these detailed experiments, Figure 1 contains the results of a computational example which illustrates our main theorem on a synthetic dataset. We postpone details of this experimental set-up to Section 4, but studying the figure at this point already gives useful insight: for fixed values of ϵ the value of $\mathcal{I}$ goes to zero at a rate proportional to γ^2 until an inflection point, around $\gamma^2 \approx \epsilon^{\min\{1,\alpha\}}$, after which the error saturates; the saturation levels themselves go to zero like $\epsilon^{\min\{1,\alpha\}}$. These facts are exactly as predicted by our theory.

The rest of this article is structured as follows. We outline the details of the Bayesian SSR problem in Section 2, introducing the likelihood and the prior in Subsections 2.1 and 2.2 followed by an analytic expression for the posterior measure in Subsection 2.3. Section 3 is dedicated to our consistency analysis and presents detailed versions of our primary results that are summarized in the Main Theorem. We first analyze the disconnected graph case in Subsection 3.1 to gain some insight into the behavior of the posterior. We then study the weakly connected graph setting in Subsection 3.2. We present the proofs of these results, relying on lemmata that are stated in Section 3, but defering their proof to Appendix A. We collect numerical experiments in Section 4 that demonstrate the sharpness of the contraction rates and bounds obtained in Section 3. We present experiments which illustrate situations in which the label noise dominates the closeness to clustering, and vice versa. We conclude the article in Section 5 with further discussion, including potential new lines of research stemming from our results. Appendix A contains the detailed proofs of the lemmata that support the main theoretical results developed in Section 3; these are also illustrated by numerical results presented in Subsections B.1 and B.2 in the supplemental material.

2. Bayesian Formulation Of SSR.

In this section we outline the Bayesian formulation of the SSR problem in detail. We derive the likelihood potential Φ in Subsection 2.1 and construct the prior measure in Subsection 2.2. An analytic expression for the posterior measure is given in Subsection 2.3.

Throughout the following we let $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product and $|\cdot|$ the Euclidean norm; we use $\|\cdot\|_2$ to denote the induced operator Euclidean norm on matrices. Recall that $\|\cdot\|_F$ denotes the Frobenius norm on matrices and define $\langle A, B\rangle_F := \mathrm{Tr}\left(A^T B\right)$, the inner-product which induces this norm. We use $\otimes$ to denote the Kronecker product between matrices. Occasionally we use $|S|$ to denote the cardinality of a set S; confusion with the Euclidean distance should not arise as we will clarify the notation based on the context.

2.1. The Likelihood.

Based on the generative model (1.1) for the labeled data $Y \in \mathbb{R}^{M\times J}$ we define the likelihood distribution $\mathbb{P}(Y|U)$ with density proportional to

$$\exp\left(-\frac{1}{2\gamma^2}\left\|UH^T - Y\right\|_F^2\right). \tag{2.1}$$

It is therefore convenient to define the likelihood potential Φ

$$\Phi : \mathbb{R}^{M\times N} \times \mathbb{R}^{M\times J} \mapsto \mathbb{R}^+, \qquad \Phi(U;Y) := \frac{1}{2}\|UH^T - Y\|_F^2. \tag{2.2}$$

Remark 2.1. We note that if the noise η is not independent but rather correlated, then the expression (2.2) needs to be modified by weighting the $\|\cdot\|_F$ norm by the inverse square root of the covariance operator of η. This will make no significant difference to what follows and we work with i.i.d. noise only to simplify the exposition.

2.2. The Prior.

We now detail the Gaussian prior measure construction and demonstrate how it expresses the geometric information in the unlabeled data X. We construct a weighted graph $G = \{Z, W\}$ with vertices Z and self-adjoint weighted adjacency matrix $W = (w_{ij})$. The weights $w_{ij} \geq 0$ reflect the affinity of data pairs $(x_i, x_j) \in X \times X$, the edge set of the graph. For example, we may construct W using a kernel $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$ by setting

$$w_{ij} = \kappa(|\mathbf{x}_i - \mathbf{x}_j|). \tag{2.3}$$

The kernel κ is assumed to be positive, non-increasing, and with bounded variance; a natural example is the Gaussian kernel $\kappa(t) = \exp\left(-|t|^2/r^2\right)$, or the indicator function of the interval $[0, r]$, both with bandwidth $r \in \mathbb{R}^+$. Note that (2.3) implies that W is symmetric and the suggested weight constructions lead to w_{ij} which encode the pairwise similarities between the points in X.

Given a weight matrix W with the properties illustrated by this explicit construction, we introduce a *graph Laplacian* operator on G of the form

$$L = D^{-p}(D - W)D^{-p}, \tag{2.4}$$

where $D = \mathrm{diag}\{d_i\}$ with entries $d_i := \sum_{j\in Z} w_{ij}$ is the diagonal degree matrix and $p \in \mathbb{R}$ is a user-defined parameter. Taking $p = 0$ gives the *unnormalized* Laplacian while $p = 1/2$ gives the *normalized* Laplacian. Other normalizations of L are also possible and can result in non-symmetric operators; see [20, Sec. 5.1] for a detailed discussion.

With the graph Laplacian matrix identified we finally define the prior covariance matrix $C_\tau \in \mathbb{R}^{N\times N}$ with hyperparameters $\tau^2, \alpha > 0$

$$C_\tau := \tau^{2\alpha}(L + \tau^2 I_N)^{-\alpha}. \tag{2.5}$$

Graph Laplacian operators are positive semi-definite (see [38, Prop. 1]); the matrix C_τ is therefore strictly positive definite thanks to the shift by $\tau^2 I_N$. The normalization by $\tau^{2\alpha}$ ensures that the largest eigenvalue of C_τ is one, while $\alpha > 0$ controls the rate of decay of the rest of the eigenvalues of C_τ; when the graph Laplacian is constructed from nearly clustered data, C_τ will exhibit a spectral gap and the eigenvectors associated with eigenvalues near one will contain geometric information about the clusters; we refer to this phenomenon as the *smoothing effect* of C_τ.

With C_τ at hand we conclude our definition of the prior on the unknown U, the Gaussian measure $\mu_0(\mathrm{d}U) = \mathcal{N}(0, I_M \otimes C_\tau)$ with Lebesgue density

$$\mu_0(\mathrm{d}U) := \frac{1}{[(2\pi)^N \det(C_\tau)]^{\frac{M}{2}}} \exp\left(-\frac{1}{2}\langle U^T, C_\tau^{-1} U^T\rangle_F\right) \mathrm{d}U. \tag{2.6}$$

If we introduce the rows $\{\mathbf{u}_1, \cdots, \mathbf{u}_M\}$ of U then we note the prior can be written as

$$\mu_0(\mathrm{d}U) = \frac{1}{[(2\pi)^N \det(C_\tau)]^{\frac{M}{2}}} \prod_{\ell=1}^{M} \exp\left(-\frac{1}{2}\left\langle \mathbf{u}_\ell, C_\tau^{-1}\mathbf{u}_\ell\right\rangle\right) \mathrm{d}\mathbf{u}_\ell.$$

The above expression reveals that, a priori, each row of U has the same distribution, and is independent of the others, and that this distribution on rows favours structure across Z which reflects the eigenvectors of the largest eigenvalues of C_τ. The matrix C_τ is chosen so that this eigenstructure reflects clustering present in the unlabeled data, for appropriately chosen τ, determined through the analysis in this paper.

Remark 2.2. The prior covariance C_τ defined in (2.5) depends on the unlabeled data X through the matrix L and the weight matrix W. This perspective differs significantly from standard BIPs, where the data only appears in the likelihood and the prior is constructed independent of the data (other than, perhaps, a noise-dependent scaling). In our formulation of SSR the labeled data appear in the likelihood potential Φ while the unlabeled data are used to construct the prior measure μ_0.

2.3. The Posterior. Using Bayes' rule we can determine the posterior μ^Y from the likelihood $\mathbb{P}(Y|U)$ and prior μ_0 defined through the Radon-Nikodym derivative

$$\frac{\mathrm{d}\mu^Y}{\mathrm{d}\mu_0}(U) = \frac{1}{\vartheta(Y)} \exp\Big(-\Phi(U;Y)\Big). \tag{2.7}$$

The posterior measure μ^Y is the Gaussian defined by

$$\mu^Y(\mathrm{d}U) = \frac{1}{\vartheta(Y)} \exp\left(-\frac{1}{2\gamma^2}\left\|UH^T - Y\right\|_F^2 - \frac{1}{2}\left\langle U^T, C_\tau^{-1}U^T\right\rangle_F\right) \mathrm{d}U. \tag{2.8}$$

It is well-known that linear inverse problems with additive Gaussian noise and a Gaussian prior result in Gaussian posteriors – this is due to the conjugacy of the prior and the likelihood. In this case we have the additional property that the independence of the rows $\mathbf{u}_\ell$ of U under the prior μ_0 is preserved under the posterior μ^Y. To see this we introduce the rows $\{\mathbf{y}_1, \cdots, \mathbf{y}_M\}$ of Y and note that we may write

$$\mu^Y(\mathrm{d}U) \propto \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\frac{1}{\gamma^2}\left|H\mathbf{u}_m - \mathbf{y}_m\right|^2 + \left\langle \mathbf{u}_m, C_\tau^{-1}\mathbf{u}_m\right\rangle\right].$$

Using this structure as the product of i.i.d. Gaussians in each of the M rows of U, Proposition A.1 shows that $\mu^Y = \mathcal{N}(U^*, I \otimes C^*)$ where $U^* \in \mathbb{R}^{M\times N}$ is the matrix with rows

$$\mathbf{u}_m^* = \frac{1}{\gamma^2}C^* H^T \mathbf{y}_m, \qquad m = 1, \ldots, M,$$

and C^* is the covariance matrix

$$C^* = \left(C_\tau^{-1} + \frac{1}{\gamma^2}H^T H\right)^{-1}.$$

3. Consistency Of Bayesian SSR.

In this section we prove consistency of the posterior μ^Y. We study consistency with respect to two small parameters: γ, which measures noise in the the labeled data Y, and ϵ which measures the closeness to perfectly clustered unlabeled data X. Recall from the Main Theorem that our goal is to show that the measure of contraction $\mathcal{I}$ (defined in (1.2)) is controlled with the noise standard deviation γ or the geometric perturbation parameter ϵ, whenever the prior hyperparameters τ, α are chosen appropriately. We will show that letting $\gamma \to 0$ results in posterior contraction, until a floor is reached that is determined by ϵ. Furthermore the analysis will reveal guidance about the choice of the hyperparameters τ and α in the prior. In Section 3.1 we consider the case of a disconnected graph with $\epsilon = 0$ and obtain contraction rates with respect to γ. In Section 3.2 we build on the disconnected case to obtain our desired results for weakly connected graphs with ϵ small.

3.1. Disconnected Graph.

Consider a weighted graph $G_0 = \{Z, W_0\}$ consisting of $K < N$ connected components $\widetilde{G}_k$, i.e., the subgraphs $\widetilde{G}_k$ are pathwise connected — any two vertices can be joined by a path within $\widetilde{G}_k$ — but the weight of edges that connect two distinct components $\widetilde{G}_i, \widetilde{G}_\ell$ are zero. Without loss of generality, we assume that the nodes in Z are ordered so that $Z = \{\widetilde{Z}_1, \widetilde{Z}_2, \cdots, \widetilde{Z}_K\}$ with the $\widetilde{Z}_k$ denoting the index set of vertices in subgraph $\widetilde{G}_k$. We refer to $\widetilde{Z}_k$ as the clusters and let $\widetilde{N}_k = |\widetilde{Z}_k|$ denote the number of vertices in the k-th cluster. We make the following assumptions on the graph G_0.

Assumption 3.1. *The graph $G_0 = \{Z, W_0\}$ satisfies the following conditions:*

(a) The weighted adjacency matrix $W_0 \in \mathbb{R}^{N\times N}$ is block diagonal

$$W_0 = \mathrm{diag}(\widetilde{W}_1, \widetilde{W}_2, \cdots, \widetilde{W}_K),$$

with $\widetilde{W}_k \in \mathbb{R}^{\widetilde{N}_k \times \widetilde{N}_k}$ denoting the weight matrices of the subgraphs $\widetilde{G}_k$.

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

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

with $\widetilde{D}_k$ denoting the degree matrix of $\widetilde{W}_k$. There exists a uniform constant $\theta > 0$ so that for $k = 1, \cdots, K$ the submatrices $\widetilde{L}_k$ satisfy

$$\langle \mathbf{x}, \widetilde{L}_k \mathbf{v}\rangle \geq \theta\langle \mathbf{v}, \mathbf{x}\rangle, \tag{3.1}$$

for all vectors $\mathbf{v} \in \mathbb{R}^{\widetilde{N}_k}$ *and* $\mathbf{v} \perp \widetilde{D}_k^p \mathbf{1}$ *with* $\mathbf{1} \in \mathbb{R}^{\widetilde{N}_k}$ *denoting the vector of ones. In other words the* $\widetilde{L}_k$ *have a uniform spectral gap.*

Remark 3.2. The existence of such θ as in (3.1) follows from [38, Props. 2 and 3], which states that 0 is an eigenvalue of $\widetilde{L}_k$ with multiplicity 1 and that the corresponding eigenvector is $\widetilde{D}_k^p \mathbf{1}$, under the pathwise connected assumption.

With a disconnected graph G_0 as above we proceed as in Section 2.2 and define graph Laplacian and covariance matrices of the form

$$L_0 := D_0^{-p}(D_0 - W_0)D_0^{-p}, \quad \text{and} \quad C_{\tau,0} := \tau^{2\alpha}(L_0 + \tau^2 I_N)^{-\alpha}, \tag{3.2}$$

with D_0 denoting the diagonal degree matrix of W_0 and parameters $\tau, \alpha > 0$. Note that

$$L_0 = \operatorname{diag}(\widetilde{L}_1, \widetilde{L}_2, \cdots, \widetilde{L}_K),$$

and that $C_{\tau,0}$ inherits a similar block-diagonal structure. We use the covariance matrix $C_{\tau,0}$ to define prior measures μ_0 of the form (2.6). In order to show posterior contraction with such a prior we also need to make some assumptions on the index set of labeled data Z' and the ground truth matrix $U^\dagger$; these encode the idea that the labels are coherent with the geometric structure implied by the perfect clustering of the data.

Assumption 3.3. *At least one label is observed in each cluster* $\widetilde{Z}_k$*; that is,*

$$|Z' \cap \widetilde{Z}_k| > 0 \qquad \forall k = 1, \ldots, K.$$

Assumption 3.4. *Let* $(\mathbf{u}_m^\dagger)^T$ *for* $m = 1, \ldots, M$ *denote the rows of* $U^\dagger$*. Then* $\mathbf{u}_m^\dagger \in \operatorname{span}\{\bar{\boldsymbol{\chi}}_1, \ldots, \bar{\boldsymbol{\chi}}_K\}$*, where the weighted set functions are defined by*

$$\bar{\boldsymbol{\chi}}_k := \frac{D_0^p \mathbf{1}_k}{|D_0^p \mathbf{1}_k|}, \tag{3.3}$$

with $\mathbf{1}_k \in \mathbb{R}^N$ *denoting indicator of the cluster* $\widetilde{Z}_k$*.*

Remark 3.5. We note here that our current exposition does not address posterior contraction when Assumption 3.4 is violated. While this is an interesting and practically pertinent question, we delay it for future study. We conjecture that as long as the ground truth variable $U^\dagger$ is consistent with the observed labeling and the true underlying clustering structure of the unlabeled data X, then posterior contraction will occur around the projection of $U^\dagger$ onto $\operatorname{span}\{\bar{\boldsymbol{\chi}}_1, \ldots, \bar{\boldsymbol{\chi}}_K\}$.

With the above assumptions in hand we are ready to present our first posterior contraction result in the case of disconnected graphs.

Theorem 3.6. *Suppose that Assumptions 3.1, 3.3 and 3.4 are satisfied in turn by the disconnected graph* G_0*, the labeled set* Z' *and the ground truth matrix* $U^\dagger$*. Consider the label model (1.1), the prior measure* $\mu_0(\mathrm{d}U) = \mathcal{N}(0, C_{\tau,0})$ *as in (2.6), and the resulting posterior measure* $\mu^Y(\mathrm{d}U)$ *as in (2.8). Then there is a constant* $\Xi > 0$ *so that, for every fixed* $\gamma, \tau, \alpha > 0$*,*

$$\mathcal{I}(\gamma, \alpha, \tau) \le \Xi \max\left\{\gamma^2, \tau^{2\alpha}\right\}\left(1 + \max\left\{\gamma^2, \tau^{2\alpha}\right\} \|U^\dagger\|_F^2\right).$$

We prove this theorem in Section 3.1.1; here we discuss the intuition behind it. If $U \sim \mu_0$ as above then $\mathbf{u}_m \overset{iid}{\sim} \mathcal{N}(0, C_{\tau,0})$ where we recall $(\mathbf{u}_m)^T$ are the rows of U.

Thus by the Karhunen-Loéve (KL) theorem,

$$\mathbf{u}_m \overset{d}{=} \sum_{j=1}^{N} \frac{1}{\sqrt{\lambda_{j,0}}} \xi_{mj} \boldsymbol{\phi}_{j,0},$$

with $\{(\lambda_{j,0}, \boldsymbol{\phi}_{j,0})\}_{j=1}^{N}$ denoting the eigenpairs of $C_{\tau,0}$ and $\xi_{mj} \overset{iid}{\sim} \mathcal{N}(0,1)$. The matrix L_0 has a K dimensional null-space spanned by the $\bar{\boldsymbol{\chi}}_k$ and this null-space is associated to the eigenvalue 1 for $C_{\tau,0}$. Furthermore, when τ^2 is small the remaining eigenvalues of $C_{\tau,0}$ are also small. These ideas are made rigorous in [21, Lemm. A.2 and Prop. A.3]. From those results it follows that

$$\mathbf{u}_m \overset{d}{=} \sum_{j=1}^{K} \xi_{mj} \bar{\boldsymbol{\chi}}_j + \mathcal{O}(\tau^{2\alpha}), \tag{3.4}$$

meaning that the prior is concentrated on $\mathrm{span}\{\bar{\boldsymbol{\chi}}_1, \ldots, \bar{\boldsymbol{\chi}}_K\}$. On the other hand the posterior μ^Y also decouples along the rows $\mathbf{u}_m$ following Proposition A.1 and so the SSR problem can be viewed as M separate BIPs for each row of $\mathbf{u}_m$, all with the same structure. As $\tau \to 0$ the prior mass concentrates on the K dimensional subspace spanned by the set-functions $\bar{\boldsymbol{\chi}}_k$. Since the posterior is absolutely continuous with respect to the prior, the posterior mass will also concentrate on the same subspace. The assumptions on the ground truth $U^\dagger$ ensure that the data is consistent with the rows $\mathbf{u}_m$ lying in this subspace and give information on assignation of labels, corresponding to weights on the $\bar{\boldsymbol{\chi}}_m$. Hence, letting $\gamma \to 0$ yields concentration of the posterior around the ground truth matrix $U^\dagger$ under Assumptions 3.3 and 3.4.

Remark 3.7. Theorem 3.6 suggests that, in this perfectly clustered setting, choosing τ to achieve $\tau^{2\alpha} = \gamma^2$ is optimal, since it balances the two sources of error in the contraction rate. However, in the next subsection we introduce a third small parameter, ϵ, measuring proximity of the unlabeled data to being perfectly clustered. We state our theorems in a setting in which τ scales as a power of ϵ, rather than γ. We make this choice because τ and ϵ are linked intrinsically through the unsupervised learning task encapsulated in the prior measure, based on the unlabeled data, whilst γ enters separately through the likelihood, which captures the labeled data. In a broader picture these considerations about choice of τ suggest the importance of choosing this hyperparameter in a data-adaptive fashion, and the importance of using hierarchical Bayesian methods to learn such parameters.

3.1.1. Proof of Theorem 3.6.

We first bound the inner expectation in (1.2), which is the mean square error of the estimator $U|Y$. We define the matrix C_0^* to denote the posterior covariance obtained by substituting the prior covariance $C_{\tau,0}$ from (3.2) into (A.2), i.e.,

$$C_0^* := \left(C_{\tau,0}^{-1} + \frac{1}{\gamma^2} B \right)^{-1}. \tag{3.5}$$

For brevity we suppress the dependence of C_0^* on τ, α, and γ. We then have

$$\mathbb{E}_{U|Y} \|U - U^\dagger\|_F^2 = \sum_{m=1}^{M} \mathbb{E}_{\mathbf{u}_m|\mathbf{y}_m} \left| \mathbf{u}_m - \mathbf{u}_m^\dagger \right|^2 = M \mathrm{Tr}(C_0^*) + \sum_{m=1}^{M} \left| \frac{1}{\gamma^2} C_0^* H^T \mathbf{y}_m - \mathbf{u}_m^\dagger \right|^2.$$

The first identity relies on the independence of the rows $\mathbf{u}_m^T$ of U under the posterior distribution, as established in Proposition A.1. The second identity comes from the fact that the mean square error is the sum of the variance and squared bias of the estimator of each row.

We may now apply the outer expectation in definition of $\mathcal{I}$ with respect to the data $Y|U^\dagger$, and since $\mathrm{Tr}(C_0^*)$ does not depend on Y, we may pull it out of the outer expectation and write

$$\mathcal{I}(\gamma, \alpha, \tau) = M\mathrm{Tr}(C_0^*) + \mathbb{E}_{Y|U^\dagger}\left(\sum_{m=1}^{M}\left|\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m - \mathbf{u}_m^\dagger\right|^2\right). \tag{3.6}$$

Since we assumed

$$\mathbf{y}_m|\mathbf{u}_m^\dagger \sim \mathcal{N}(H\mathbf{u}_m^\dagger, \gamma^2 I_J) \tag{3.7}$$

and the rows $\{\mathbf{y}_m^T\}_{m=1}^M$ are independent conditional on $U^\dagger$, we can write

$$\mathbb{E}_{Y|U^\dagger}\left|\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m - \mathbf{u}_m^\dagger\right|^2 = \mathbb{E}_{\mathbf{y}_m|\mathbf{u}_m^\dagger}\left|\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m - \mathbf{u}_m^\dagger\right|^2.$$

This expectation is the mean square error of the posterior mean estimator of $\mathbf{u}_m^\dagger$, which can be decomposed once more into a variance and a squared bias term:

$$\mathbb{E}_{\mathbf{y}_m|\mathbf{u}_m^\dagger}\left|\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m - \mathbf{u}_m^\dagger\right|^2 = \mathrm{Tr}\left(\mathrm{Cov}\left(\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m\right)\right) + \left|\mathbb{E}_{\mathbf{y}_m|\mathbf{u}_m^\dagger}\left(\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m\right) - \mathbf{u}_m^\dagger\right|^2,$$

where $\mathrm{Cov}(\cdot)$ denotes the covariance matrix of a random vector. We compute the variance term using (3.7) once more:

$$\mathrm{Cov}\left(\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m\right) = \frac{1}{\gamma^2}C_0^* H^T \mathrm{Cov}\left(\mathbf{y}_m\right)\frac{1}{\gamma^2}H(C_0^*)^T = \frac{1}{\gamma^2}C_0^* B C_0^*,$$

where we used the fact that $\mathrm{Cov}(\mathbf{y}_m) = \gamma^2 I_J$ and $B = H^T H \in \mathbb{R}^{N\times N}$. As for the bias term, we can write

$$\mathbb{E}_{\mathbf{y}_m|\mathbf{u}_m^\dagger}\left(\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m\right) = \frac{1}{\gamma^2}C_0^* H^T H\mathbf{u}_m^\dagger = \frac{1}{\gamma^2}C_0^* B\mathbf{u}_m^\dagger.$$

Putting these terms together yields

$$\mathbb{E}_{\mathbf{y}_m|\mathbf{u}_m^\dagger}\left|\frac{1}{\gamma^2}C_0^* H^T \mathbf{y}_m - \mathbf{u}_m^\dagger\right|^2 = \frac{1}{\gamma^2}\mathrm{Tr}\left(C_0^* B C_0^*\right) + \left|\frac{1}{\gamma^2}C_0^* B\mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger\right|^2.$$

Substituting this identity back into (3.6) yields

$$\mathcal{I}(\gamma, \alpha, \tau) = M\mathrm{Tr}(C_0^*) + \frac{M}{\gamma^2}\mathrm{Tr}(C_0^* B C_0^*) + \sum_{m=1}^{M}\left|\frac{1}{\gamma^2}C_0^* B\mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger\right|^2. \tag{3.8}$$

The desired bound now follows from Lemmata 3.8, 3.9, and 3.10 below that in turn bound the first, second, and third term in the right hand side of (3.8). These Lemmata are proved in Appendix A.2.

LEMMA 3.8. *Suppose Assumptions 3.1 and 3.3 are satisfied by the disconnected graph G_0 and the labeled set Z' respectively. Then there exists a constant $\Xi > 0$, such that for any $\gamma, \tau, \alpha > 0$,*

$$\operatorname{Tr}(C_0^*) \leq \Xi \max\{\gamma^2, \tau^{2\alpha}\}, \tag{3.9}$$

where C_0^* is the posterior covariance matrix in (3.5).

LEMMA 3.9. *Suppose Lemma 3.8 is satisfied. Then for any $\gamma, \tau, \alpha > 0$*

$$\frac{1}{\gamma^2}\operatorname{Tr}(C_0^* B C_0^*) \leq \Xi \max\left\{\gamma^2, \tau^{2\alpha}\right\},$$

with the same constant $\Xi > 0$ as in (3.9).

LEMMA 3.10. *Suppose Assumptions 3.1, 3.3, and 3.4 are in turn satisfied by the disconnected graph G_0, the labeled set Z', and the ground truth function $U^\dagger$. Then for any $\gamma, \tau, \alpha > 0$ and $m = 1, \ldots, M$,*

$$\left|\frac{1}{\gamma^2} C_0^* B \mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger\right| \leq \Xi \max\{\gamma^2, \tau^{2\alpha}\},$$

where $\Xi > 0$ is the same constant as in (3.9).

3.2. Weakly Connected Graph. We now consider a generalization of the setting in the previous subsection, in which the disconnected graph $G_0 = \{Z, W_0\}$ is perturbed, and the perturbation results in a connected graph $G_\epsilon = \{Z, W_\epsilon\}$. Following [21] we collect the following set of assumptions on this perturbed graph G_ϵ.

ASSUMPTION 3.11. *The graph $G_\epsilon = \{Z, W_\epsilon\}$ satisfies the following three conditions.*

(a) The weight matrix W_ϵ can be expanded in the form

$$W_\epsilon = W_0 + \sum_{h=1}^{\infty} \epsilon^h W^{(h)}, \tag{3.10}$$

where W_0 is the weighted adjacency matrix of a disconnected graph G_0.

(b) The matrices $W^{(h)}$ are self-adjoint and $\{\|W^{(h)}\|_2\}_{h=1}^\infty \in \ell^\infty$.

(c) Let $w_{ij}^{(0)}$ and $w_{ij}^{(h)}$ denote the entries of W_0 and $W^{(h)}$ respectively. Then, for $h \geq 1$,

$$\begin{cases} w_{ij}^{(h)} \geq 0, & \text{if} \quad w_{ij}^{(0)} = 0 \quad \text{for} \quad i, j \in Z, i \neq j. \\ w_{ii}^{(h)} = 0. \end{cases} \tag{3.11}$$

The assumptions (b) and (c) above ensure that W_ϵ is a well-defined adjacency matrix. Also note that (c) allows for $w_{ij}^{(h)}$, $h \geq 1$, to be negative whenever $w_{ij}^{(0)} > 0$. With the above assumptions identified we can proceed analogously to Section 2.2 to define Laplacian and covariance matrices

$$L_\epsilon := D_\epsilon^{-p}(D_\epsilon - W_\epsilon)D_\epsilon^{-p}, \quad \text{and} \quad C_{\tau,\epsilon} := \tau^{2\alpha}(L_\epsilon + \tau^2 I_N)^{-\alpha}, \tag{3.12}$$

with D_ϵ denoting the diagonal degree matrix of W_ϵ and parameters $\tau, \alpha > 0$. We then use the covariance matrix $C_{\tau,\epsilon}$ to define a prior measure μ_0 of the form (2.6) on the weakly connected graph G_ϵ. With the assumptions made about the disconnected set-up in Subsection 3.1, and the above new assumptions on the weakly connected set-up, we can now present our main posterior contraction result, the analogue of Theorem 3.6, for weakly connected graphs G_ϵ.

THEOREM 3.12. *Suppose Assumptions 3.1, 3.3, 3.4 and 3.11 are satisfied in turn by the disconnected graph G_0, the labeled set Z', the ground truth matrix $U^\dagger$ and the weakly connected graph G_ϵ. Fix $\alpha > 0$. Then there exist constants $\epsilon_0 \in (0,1)$ and $\Xi, \Xi_1 > 0$ such that the following holds uniformly for any sequence $(\epsilon, \tau, \gamma) \to 0$, along which $\epsilon \leq \epsilon_0 \tau^2$:*

$$\mathcal{I}(\gamma, \alpha, \tau, \epsilon) \leq \Xi \max\left\{\gamma^2, \left(\frac{\tau^2}{1 - \Xi_1 \epsilon/\tau^2}\right)^\alpha\right\} \times \left(1 + \max\left\{\gamma^2, \left(\frac{\tau^2}{1 - \Xi_1 \epsilon/\tau^2}\right)^\alpha\right\}\left[\epsilon + \frac{\epsilon}{\tau^{2\alpha}} + \left(1 + \frac{\epsilon}{\tau^2}\right)^\alpha\right]^2 \|U^\dagger\|^2\right).$$

The intuition behind the proof is that we use the same ideas which underlie Theorem 3.6, which concerns the case $\epsilon = 0$, coupled with new arguments which control perturbations to the spectrum of $C_{\tau,\epsilon}$ with respect to that of $C_{\tau,0}$. Specifically $C_{\tau,\epsilon}$ now has a one-dimensional null-space associated with the eigenvalue 1, but has an additional $K-1$ eigenvalues of size $1 - \mathcal{O}(\epsilon/\tau^2)$. The remaining eigenvalues are small, of $\mathcal{O}(\tau^{2\alpha})$, if an appropriate relationship between ϵ and τ is imposed. The eigenfunctions associated with the K eigenvalues at, or near, 1, nearly span the same space as the N weighted set-functions $\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$. Let $(\mathbf{u}_m)^T$ denote the rows of $U \sim \mu_0$. Then it follows from [21, A.10] that these rows concentrate on the span of the $\bar{\boldsymbol{\chi}}_k$ with errors of the form $\mathcal{O}\left(\epsilon^2\tau^{-4} + \tau^{4\alpha} + \epsilon^2\right)$ when $\epsilon = o(\tau^2)$ and τ^2 is small and of the form $\mathcal{O}\left(\tau^{4\alpha} + \epsilon^2\right)$ when $\epsilon = \Theta(\tau^2)$ and τ^2 is small. These approximation results for the rows $(\mathbf{u}_m)^T$ under the prior underlie the proof. The rest of the proof follows in the footsteps of Theorem 3.6. First, we decouple the posterior on the rows of U using Proposition A.1 to obtain M independent BIPs. In each BIP the prior concentration on the span of $\bar{\boldsymbol{\chi}}_k$ results in posterior concentration along the same subspace, at which point, the noise standard deviation γ in the likelihood potential Φ controls the contraction of the posterior around the ground truth matrix $U^\dagger$ under Assumptions 3.3 and 3.4.

3.2.1. Proof of Theorem 3.12. Let us define the perturbed posterior covariance matrix

$$C_\epsilon^* := \left(C_{\tau,\epsilon}^{-1} + \frac{1}{\gamma^2}B\right)^{-1}, \tag{3.13}$$

following (3.12) with the prior covariance matrix $C_{\tau,\epsilon}$. Observe that the arguments leading up to the upper bound (3.1.1) hold with C_0^* replaced with C_ϵ^*. Thus we immediately obtain the identity

$$\mathcal{I}(\gamma, \alpha, \tau, \epsilon) = M\mathrm{Tr}(C_\epsilon^*) + \frac{M}{\gamma^2}\mathrm{Tr}(C_\epsilon^* B C_\epsilon^*) + \sum_{m=1}^M \left|\frac{1}{\gamma^2}C_\epsilon^* B \mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger\right|^2. \tag{3.14}$$

Similarly to Section 3.1.1 we prove Theorem 3.12 by bounding each term in the right hand side of (3.14) in the Lemmata 3.13, 3.14, and 3.15 below. The proofs are collected in Appendix A.3.

LEMMA 3.13. *Suppose Assumptions 3.1, 3.3, and 3.11 are satisfied in turn by the disconnected graph G_0, the labeled set Z', and the weakly connected graph G_ϵ. Fix $\alpha > 0$. Then there exist constants $\epsilon_0 \in (0,1)$ and $\Xi_0, \Xi_1 > 0$ such that the following holds uniformly along any sequence $(\epsilon, \tau, \gamma) \to 0$, along which $\epsilon \leq \epsilon_0\tau^2$:*

$$\mathrm{Tr}(C_\epsilon^*) \leq \Xi_0 \max\left\{\gamma^2, \left(\frac{\tau^2}{1 - \Xi_1\epsilon/\tau^2}\right)^\alpha\right\},$$

with C_ϵ^* as in (3.13).

LEMMA 3.14. *Suppose that the conditions of Lemma 3.13 are satisfied and fix $\alpha > 0$. Then there exist constants $\epsilon_0 \in (0,1)$ and $\Xi_0, \Xi_1 > 0$ such that the following holds uniformly for any sequence $(\epsilon, \tau, \gamma) \to 0$, along which $\epsilon \leq \epsilon_0 \tau^2$:*

$$\frac{1}{\gamma^2}\mathrm{Tr}(C_\epsilon^* B C_\epsilon^*) \leq \Xi_0 \max\left\{\gamma^2, \left(\frac{\tau^2}{1-\Xi_1\epsilon/\tau^2}\right)^\alpha\right\},$$

where $\epsilon_0 \in (0,1)$ and $\Xi_0, \Xi_1 > 0$ are the same constants as in Lemma 3.13.

LEMMA 3.15. *Suppose Assumptions 3.1, 3.3, 3.4, and 3.11 are satisfied by the disconnected graph G_0, the labeled set Z', the ground truth matrix $U^\dagger$ and the weakly connected graph G_ϵ respectively and fix $\alpha > 0$. Then there exist constants $\epsilon_0 \in (0,1)$ and $\Xi_1, \Xi_2 > 0$ such that the following holds uniformly for any sequence $(\epsilon, \tau, \gamma) \to 0$, along which $\epsilon \leq \epsilon_0 \tau^2$:*

$$\left|\frac{1}{\gamma^2}C_\epsilon^* B \mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger\right| \leq \Xi_2 \max\left\{\gamma^2, \left(\frac{\tau^2}{1-\Xi_1\epsilon/\tau^2}\right)^\alpha\right\}\left[\epsilon + \frac{\epsilon}{\tau^{2\alpha}} + \left(1+\frac{\epsilon}{\tau^2}\right)^\alpha\right]|\mathbf{u}_m^\dagger|.$$

3.3. Corollary - Main Theorem.

We now present a corollary of Theorem 3.12 that is the precisely stated version of our informal Main Theorem from Section 1.

COROLLARY 3.16. *Suppose that the conditions of Theorem 3.12 are satisfied and that for a fixed $\alpha > 0$, the hyperparameters (ϵ, τ) are chosen to satisfy*

$$2\Xi_1 \epsilon = \tau^{\max\{2, 2\alpha\}}.$$

Then there exists $\Xi_2 > 0$ depending on α and the constants Ξ, Ξ_1 from Theorem 3.12 but independent of ϵ and γ, so that

$$\mathcal{I} \leq \Xi_2 \max\left\{\gamma^2, \epsilon^{\min\{1,\alpha\}}\right\}.$$

Remark 3.17. The reader is encouraged to study the discussion following the informal Main Theorem for an interpretation of this result in terms of asymptotic consistency. We also note that an application of Markov's inequality can immediately extend the bound in Corollary 3.16 to a bound on the expected probabilities of posterior samples being found far from the ground truth $U^\dagger$. More precisely we have that for any $\delta > 0$

$$\mathbb{E}_{Y|U^\dagger}\left\{\mu^Y\left(\left\|U - U^\dagger\right\|_F > \delta\right)\right\} \leq \frac{\mathcal{I}}{\delta^2}.$$

4. Numerical Experiments.

In this section, we provide numerical experiments that elucidate our main theoretical results and in particular examine the convergence rate of the contraction functional $\mathcal{I}$ with respect to both the ϵ and γ parameters. We use both a synthetic example in Subsection 4.1 as well as the MNIST database of handwritten digits [25], in Subsection 4.2. In both examples we compute $\mathcal{I}$ via the decomposition given in (3.8), which provides us with an explicit formula to numerically compute the contraction measure. We then vary ϵ and γ parameters while choosing $\tau = \epsilon^{1/\max\{2,2\alpha\}}$. We numerically differentiate $\log(\mathcal{I})$ with respect to $\log(\epsilon)$ and $\log(\gamma)$ to estimate the rate of convergence with respect to these two parameters. A surface plot of these derivatives is then presented in Figures 2 and 3, for the two respective datasets, in which the color encodes the estimated rate of convergence in terms of the respective variables. The dark blue colors in these plots indicate a rate of convergence of $\mathcal{I}$ that is close to zero, meaning that convergence has ceased, while bright yellow colors indicate larger convergence rates of $\mathcal{I}$. Further numerical results are presented in Subsections B.1 and B.2 in the supplemental material, taking a closer look at the rates of convergence of different bias and variance terms that contribute to $\mathcal{I}$.

4.1. Synthetic Data. We construct a synthetic weakly connected graph consisting of three clusters of 100 nodes each, where each cluster represents a different class. We obtain the weight matrix W_ϵ following (3.10); we truncate the expansion at the ϵ^3 level. Each entry of weight matrices W_0 and $W^{(h)}$, $h = 1, 2, 3$ are drawn independently from a uniform distribution on $[0, 1]$. The matrices W_0 and $W^{(h)}$, $h = 1, 2, 3$ are fixed once sampled and are used to construct W_ϵ for different ϵ values. Each W_ϵ is then symmetrized via the transformation $W_\epsilon \mapsto (W_\epsilon + W_\epsilon^T)/2$. We pick one node from each cluster to be labeled and choose ground truth $U^\dagger = [\bar{\boldsymbol{\chi}}_1, \bar{\boldsymbol{\chi}}_2, \bar{\boldsymbol{\chi}}_3]^T$. We vary ϵ values from 10^{-1} to 10^{-15} and γ ranging from 10^{-1} to $10^{-7.5}$; τ is taken to be $\epsilon^{1/\max\{2,2\alpha\}}$.

In Figure 1, we demonstrate the convergence of $\mathcal{I}$ in the limit of the noise standard deviation γ going to zero, for different values of α and ϵ. We see posterior contraction with respect to γ until a floor is reached; this floor depends on ϵ, the degree of clustering in the data, and is smaller for smaller ϵ.

In Figure 2 we study this phenomenon in more detail. Let us define

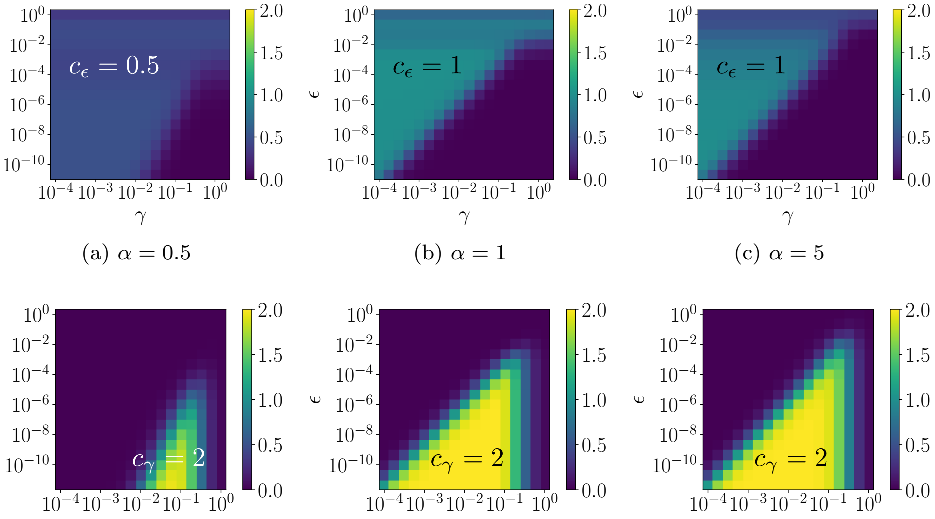
$$c_\epsilon := \partial \log(\mathcal{I})/\partial \log(\epsilon) \geq 0 \text{ and } c_\gamma := \partial \log(\mathcal{I})/\partial \log(\gamma) \geq 0,$$

which correspond to the contraction rates of $\mathcal{O}(\epsilon^{c_\epsilon})$ and $\mathcal{O}(\gamma^{c_\gamma})$ respectively. We present surface plots in Figure 2 of c_ϵ (top row) and c_γ (bottom row) as functions of ϵ, γ for various values of α. Darker (lighter) regions correspond to smaller (larger) values of the logarithmic slopes c_ϵ, c_γ. In regions with lighter values (i.e. $c_\epsilon, c_\gamma > 0$), we observe posterior contraction because the logarithmic slopes are nonzero. The darker regions correspond to instances where the contraction has ceased as indicated by the logarithmic slopes being zero. This is the phenomenon that is displayed in Figure 1, where the value of $\mathcal{I}$ reduces with respect to γ up to the point where the errors saturate at an ϵ-dependent value as predicted by the bounds in Theorem 3.12.

In the bottom row of Figure 2, horizontal "slices" of the plot correspond to a fixed value of ϵ which is how Figure 1 can be obtained. Going from right to left, we observe that the contraction rate is on the order of γ^2, until the point that $\gamma^2 \approx \epsilon^{\min\{1,\alpha\}}$ when our theory predicts that the $\mathcal{I}$ will saturate and contraction has stopped, i.e., $c = 0$. These plots illustrate the sharpness of our theoretical bounds of Theorem 3.12 for the posterior contraction measure $\mathcal{I}$. Similar results, with the roles of ϵ and γ swapped, are seen in the top row of Figure 2.

4.2. MNIST Data. In this subsection we use the MNIST dataset [25] to test our theory on a realistic dataset. MNIST is a data set of 70,000 grayscale 28×28 pixel images of handwritten digits (0-9), of which we use only the digits 1, 4, and 7. Each image is represented by a vector $\mathbf{x}_i \in \mathbb{R}^{784}$ and we normalize the pixel values to range from 0 to 1. To confirm our theory in practice presents the issue of determining how to control the parameter ϵ that is inherent to the clustering structure of a given fixed unlabeled data set X given in application. However, in this example, we may use the fact that every image is labeled and so the clustering structure of the dataset is known. Using this we may devise an ϵ-dependent parameter set to observe what happens in the $\epsilon \to 0$ limit.

First, we create a similarity graph G based on the unlabeled data X of reshaped images $\mathbf{x}_i \in \mathbb{R}^{784}$. Given the known clustering (i.e. class memberships) of the points in



(a) $\alpha = 0.5$

(b) $\alpha = 1$

(c) $\alpha = 5$

(d) $\alpha = 0.5$

(e) $\alpha = 1$

(f) $\alpha = 5$

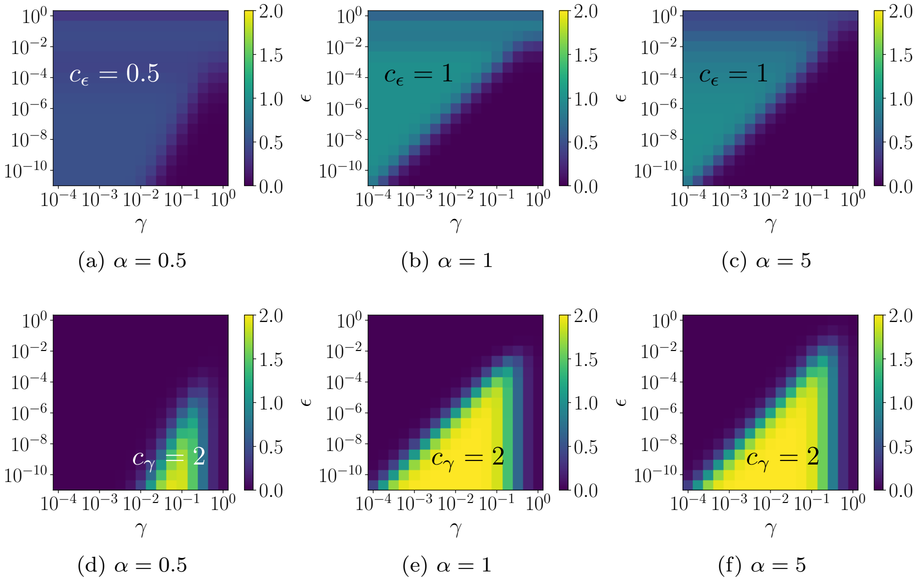
FIG. 2. *A numerical demonstration of the Main Theorem on the synthetic dataset. The top panels showcase numerical estimates of $c_\epsilon = \partial \log(\mathcal{I})/\partial \log(\epsilon)$ for different α values and the bottom panels showcase the numerical estimates of $c_\gamma = \partial \log(\mathcal{I})/\partial \log(\gamma)$. In the dark blue regions $c_\epsilon, c_\gamma \approx 0$, indicating that $\mathcal{I}$ stays flat with respect to the respective variable ϵ or γ and so contraction has ceased; the slope of the brighter regions is denoted on each figure and implies posterior contraction. The transition between the dark and bright regions occurs approximately at $\epsilon = \gamma^{2/\min\{1,\alpha\}}$.*

the MNIST dataset, we can identify the *inter-cluster* edges, those edges that connect nodes of different clusters corresponding to different digits. If the original weight matrix is given by W, with entries w_{ij}, then we scale the inter-cluster edges by ϵ to obtain W_ϵ as:

$$[W_\epsilon]_{ij} = \begin{cases} w_{ij} & \text{if } i, j \in \tilde{Z}_k, \\ \epsilon w_{ij} & \text{if } i \in \tilde{Z}_k, j \in \tilde{Z}_\ell, \text{ with } k \neq \ell. \end{cases}$$

Sending $\epsilon \to 0$ then results in a disconnected graph, where each cluster represents a different digit. For all ϵ sufficiently small the graph Laplacian will have the structure underlying our theory.

For our experiment, we sample 100 images uniformly at random from digits 1, 4, and 7. The similarity graph $W = (w_{ij})$ is constructed via the Gaussian kernel and the Zelnik-Perona scaling [42], $w_{ij} = \exp(-|\mathbf{x}_i - \mathbf{x}_j|^2/r_i r_j)$, where r_i is the Euclidean distance between data point i and its 15th nearest neighbor. Following the same procedure as the synthetic data, we pick one node from each digit to be labeled and choose the ground truth $U^\dagger = [\bar{\boldsymbol{\chi}}_1, \bar{\boldsymbol{\chi}}_2, \bar{\boldsymbol{\chi}}_3]^T$. We evaluate the contraction measurement $\mathcal{I}$ for a range of ϵ and γ. We present the results in Figure 3. It is clear that Figure 3 is nearly identical to Figure 2, demonstrating that the behaviour on this MNIST data set is close to that observed in the synthetic case; in turn the two sets of experiments together attest to the sharpness of our contraction rate estimates in Theorem 3.12. Working with the MNIST dataset highlights the relevance of our analysis to real-world SSR applications.



(a) $\alpha = 0.5$

(b) $\alpha = 1$

(c) $\alpha = 5$

(d) $\alpha = 0.5$

(e) $\alpha = 1$

(f) $\alpha = 5$

Fig. 3. *A numerical demonstration of the Main Theorem on the MNIST dataset with digits 1, 4, and 7. The top panels showcase numerical estimates of $c_\epsilon = \partial \log(\mathcal{I})/\partial \log(\epsilon)$ for different α values and the bottom panels showcase the numerical estimates of $c_\gamma = \partial \log(\mathcal{I})/\partial \log(\gamma)$. In the dark blue regions $c_\epsilon, c_\gamma \approx 0$, indicating that $\mathcal{I}$ stays flat with respect to the respective variable ϵ or γ and so contraction has ceased; the slope of the brighter regions is denoted on each figure and implies posterior contraction. The transition between the dark and bright regions occurs approximately at $\epsilon = \gamma^{2/\min\{1,\alpha\}}$. These results are strikingly similar to our synthetic experiment depicted in Figure 2.*

5. Conclusions.

The work in this paper is, to the best of our knowledge, the first analysis of Bayesian posterior consistency in semi-supervised regression (SSR). The regression formulation of semi-supervised learning is convenient for both computations and analysis due to conjugacy of Gaussian likelihoods and priors, leading to a Gaussian posterior. The resulting closed form is useful in practice [43] and for theory, such as that developed in this paper. We formulate the SSR problem as a Bayesian inverse problem in which the unlabeled data defines the prior and the labeled data defines the likelihood. By postulating coherence between the labeled and unlabeled data we are able to quantify the convergence of the posterior distribution to the truth in terms of the noise in the labels and a measure of clustering in the data. As a by-product of the analysis we also learn about parameter choices within the data-informed prior construction.

However the SSR formulation has some undesirable model characteristics relating to the fact that the latent variable U, which is real-valued, and the labels, which are categorical, are seen as elements of the same space. A fruitful avenue for future study is to combine the work in this paper with that developed in [21], where consistency of probit-based optimization is studied, in order to analyze Bayesian posterior consistency for probit-based approaches to SSL. The probit methodology postulates a link function connecting the latent variable to labels, a concrete example being the use of the sign function in binary classification [6]. Another interesting direction for theoretical analyses of SSR concerns active learning as pioneered in [45]. The framework and methodology developed here will be useful in developing principled theories of active learning.

Appendix A. Proof of Lemmata.

In this appendix we start by discussing useful properties of the posterior measure in Subsection A.1; in particular we show that the posterior is Gaussian and give closed form expressions for its mean and covariance. Subsections A.2, A.3 we present the detailed proofs of the lemmata used to prove our main results, Theorems 3.6 and 3.12. Numerical experiments which illustrate these lemmata are contained in Subsections B.1 and B.2 of the supplemental material document.

A.1. Characterizing the Posterior. Here we collect some results that completely characterize the posterior measure μ^Y as a Gaussian measure with explicit formulae for its mean and covariance.

PROPOSITION A.1. *Consider the posterior measure μ^Y given by (2.8). Then*

(i) $\mu^Y = \mathcal{N}(U^*, I_M \otimes C^*)$ *and has Lebesgue density*

$$\begin{aligned}\mu^Y(\mathrm{d}U) &= \frac{1}{\vartheta(Y)} \exp\left(-\frac{1}{2}\left\langle (U-U^*)^T, (C^*)^{-1}(U-U^*)^T\right\rangle_F\right)\mathrm{d}U \\ &\equiv \frac{1}{\vartheta(Y)} \prod_{m=1}^{M} \exp\left(-\frac{1}{2}\langle(\mathbf{u}_m - \mathbf{u}_m^*), (C^*)^{-1}(\mathbf{u}_m - \mathbf{u}_m^*)\rangle\right)\mathrm{d}\mathbf{u}_\ell.\end{aligned} \tag{A.1}$$

Here U^* is the posterior mean with rows $(\mathbf{u}_m^*)^T$ and C^* is the covariance matrix of each row $(\mathbf{u}_m^*)^T$.

(ii) The posterior means $\mathbf{u}_m^*$ and covariances C^* are given by

$$\mathbf{u}_m^* = \frac{1}{\gamma^2}C^* H^T \mathbf{y}_m, \qquad C^* = \left(C_\tau^{-1} + \frac{1}{\gamma^2}B\right)^{-1}, \tag{A.2}$$

where $B = H^T H$ and $\mathbf{y}_m^T$ are the rows of Y.

(iii) The rows $\mathbf{u}_m^T$ of $U \sim \mu^Y$ are i.i.d. according to the Gaussian distribution $\mathcal{N}(\mathbf{u}_\ell^*, C^*)$.

Proof. To show (i) we begin by expressing the likelihood in terms of the rows of U and Y,

$$\exp\left(-\Phi(U;Y)\right) = \exp\left(-\frac{1}{2\gamma^2}\left\|HU^T - Y^T\right\|_F^2\right) = \exp\left(-\frac{1}{2\gamma^2}\sum_{m=1}^{M}|H\mathbf{u}_m - \mathbf{y}_m|^2\right).$$

Combining with (2.6) we can express the Lebesgue density of the posterior as

$$\begin{aligned}\mu^Y(\mathrm{d}U) &\propto \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\left\langle \mathbf{u}_m, C_\tau^{-1}\mathbf{u}_m\right\rangle + \frac{1}{\gamma^2}\left|H\mathbf{u}_m - \mathbf{y}_m\right|^2\right] \\ &= \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\left\langle \mathbf{u}_m, C_\tau^{-1}\mathbf{u}_m\right\rangle + \frac{1}{\gamma^2}\left(\langle\mathbf{u}_m, B\mathbf{u}_m\rangle - 2\langle\mathbf{u}_m, H^T\mathbf{y}_m\rangle + |\mathbf{y}_m|^2\right)\right] \\ &\propto \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\left\langle \mathbf{u}_m, (C^*)^{-1}\mathbf{u}_m\right\rangle - 2\left\langle \mathbf{u}_m, \frac{1}{\gamma^2}H^T\mathbf{y}_m\right\rangle + \left\langle \mathbf{u}_m^*, (C^*)^{-1}\mathbf{u}_m^*\right\rangle\right] \\ &= \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\left\langle \mathbf{u}_m, (C^*)^{-1}\mathbf{u}_m\right\rangle - 2\left\langle \mathbf{u}_m, (C^*)^{-1}\mathbf{u}_m^*\right\rangle + \left\langle \mathbf{u}_m^*, (C^*)^{-1}\mathbf{u}_m^*\right\rangle\right]\end{aligned}$$

$$= \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\left\langle \mathbf{u}_m - \mathbf{u}_m^*, (C^*)^{-1}(\mathbf{u}_m - \mathbf{u}_m^*)\right\rangle\right]$$
$$= \exp\left[-\frac{1}{2}\left\langle (U-U^*)^T, (C^*)^{-1}(U-U^*)^T\right\rangle_F\right],$$

with $\mathbf{u}_m^*$, and C^* as in (A.2). Assertion (ii) follows from (A.1), and the observation that the negative log posterior is a sum of identical positive-definite quadratic forms in each $\mathbf{u}_m$, from which the expressions for mean and variance of $\mathbf{u}_m$ may be inferred. Assertion (iii) is a consequence of the fact that uncorrelated Gaussian random variables are also independent. □

A.2. Proofs of Lemmata 3.8, 3.9, and 3.10.

A.2.1. Proof of Lemma 3.8.

Proof. Let $P_0 \in \mathbb{R}^{N\times N}$ denote the projection matrix onto $\operatorname{span}\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$ (recall (3.3)) and define

$$\beta = \sqrt{\frac{K}{K+\zeta^2/4}}, \qquad \zeta := \min_{k\le K}\min_{i\in Z_k}|\bar{\boldsymbol{\chi}}_k(i)|. \tag{A.3}$$

Our method of proof is to obtain lower bounds on the Dirichlet energy $\left\langle \mathbf{v}, (C_0^*)^{-1}\mathbf{v}\right\rangle$ for unit vectors $\mathbf{v}\in\mathbb{R}^N$ by considering two cases where $|P_0\mathbf{v}|\ge\beta$ of $|P_0\mathbf{v}|<\beta$. This translates to a lower bound on the smallest eigenvalue of $(C_0^*)^{-1}$. Since $\operatorname{Tr}(C_0^*) = \sum_{j=1}^N \lambda_{j,0}$, with $\lambda_{j,0}$ denoting the strictly positive eigenvalues of C_0^*, the lower bound on the Dirichlet energy of $(C_0^*)^{-1}$ translates to an upper bound on $\operatorname{Tr}(C_0^*)$.

Case 1 ($|P_0\mathbf{v}|\ge\beta$): Since $\mathbf{v}$ is a unit vector, we have that $\|(I-P_0)\mathbf{v}\|_\infty \le |(I-P_0)\mathbf{v}|\le\sqrt{1-\beta^2}$. The matrix $C_{\tau,0}$ and its inverse are positive definite, and so

$$\left\langle \mathbf{v}, (C_0^*)^{-1}\mathbf{v}\right\rangle = \left\langle \mathbf{v}, \left(\frac{1}{\gamma^2}B\mathbf{v} + C_{\tau,0}^{-1}\right)\mathbf{v}\right\rangle \ge \left\langle \mathbf{v}, \frac{1}{\gamma^2}B\mathbf{v}\right\rangle = \frac{1}{\gamma^2}\sum_{i\in Z'} v_i^2, \tag{A.4}$$

where we used v_i to denote the entries of $\mathbf{v}$. Let us write $P_0\mathbf{v} = \sum_{k=1}^K c_k\bar{\boldsymbol{\chi}}_k$ with $c_k := \langle \mathbf{v}, \bar{\boldsymbol{\chi}}_k\rangle$ denoting the basis coefficients of $\mathbf{v}$ in span of $\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$ and define

$$\mathfrak{k} := \arg\max_k |c_k|,$$

the index of the absolutely maximal coefficient amongst the c_k. The assumption $|P_0\mathbf{v}|\ge\beta$ implies $\sum_{k=1}^K c_k^2 \ge \beta^2$. It then follows that

$$K\max_{k\le K} c_k^2 \ge \sum_{k=1}^K c_k^2 \ge \beta^2,$$

hence $|c_{\mathfrak{k}}| = \max_{k\le K}|c_k| \ge \beta/\sqrt{K}$. Since each $\bar{\boldsymbol{\chi}}_k$ is supported on $\widetilde{Z}_k$ on which it takes values that are at least ζ, we have

$$|(P_0\mathbf{v})_i| = |c_{\mathfrak{k}}|(\bar{\boldsymbol{\chi}}_{\mathfrak{k}})_i \ge \frac{\beta\zeta}{\sqrt{K}} \qquad \text{for} \qquad i\in\widetilde{Z}_{\mathfrak{k}},$$

where we used $(P_0\mathbf{v})_i$ to denote the i-th entry of the vector $P_0\mathbf{v}$. It then follows that for $i \in \widetilde{Z}_{\mathfrak{k}}$

$$\begin{aligned}|v_i| &= |(P_0\mathbf{v})_i + ((I-P_0)\mathbf{v})_i| \geq \max\{0, |(P_0\mathbf{v})_i| - \|(I-P_0)\mathbf{v}\|_\infty\} \\ &\geq \max\left\{0, \frac{\beta\zeta}{\sqrt{K}} - \sqrt{1-\beta^2}\right\}.\end{aligned}$$

Substituting the value of β from (A.3), we obtain $|v_i| \geq \left(4K/\zeta^2+1\right)^{-1/2}$. Following Assumption 3.3, i.e. $\widetilde{Z}'_k \neq \emptyset$ for all k, we have

$$\frac{1}{\gamma^2}\sum_{j\in Z'} v_j^2 \geq \frac{1}{\gamma^2}|v_i|^2 \geq \gamma^{-2}\left(4K/\zeta^2+1\right)^{-1} \qquad \text{for some index } i \in \widetilde{Z}'_{\mathfrak{k}}.$$

Putting this lower bound together with (A.4) we conclude that for any $\mathbf{v}$ such that $|P_0\mathbf{v}| \geq \beta$,

$$\langle \mathbf{v}, (C_0^*)^{-1}\mathbf{v}\rangle \geq \gamma^{-2}\left(4K/\zeta^2+1\right)^{-1}.$$

Case 2 ($|P_0\mathbf{v}| < \beta$): We naturally have $|(I-P_0)\mathbf{v}| \geq \sqrt{1-\beta^2}$. Let $\{(\sigma_{k,0}, \boldsymbol{\phi}_{k,0})\}_{k=1}^N$ denote the eigenpairs of L_0, indexed by order of increasing eigenvalues. Recall from Subsection 3.1 that $\sigma_{k,0} = 0$ for $k = 1, 2, \ldots, K$ and $\{\boldsymbol{\phi}_{k,0}\}_{k=1}^K \subset \operatorname{span}\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$. Moreover, the orthonormal eigenvectors $\{\boldsymbol{\phi}_{k,0}\}_{k=1}^N$ are also eigenvectors of $C_{\tau,0}^{-1}$. With some abuse of notation we define $c_k := \langle \mathbf{v}, \boldsymbol{\phi}_{k,0}\rangle$ for $k = K+1, \ldots, N$ and write $(I-P_0)\mathbf{v} = \sum_{k=K+1}^N c_k\boldsymbol{\phi}_{k,0}$. In light of this identity we compute

$$\text{(A.5)} \quad \begin{aligned}\langle \mathbf{v}, (C_0^*)^{-1}\mathbf{v}\rangle &= \left\langle \mathbf{x}, \left(\frac{1}{\gamma^2}B + C_{\tau,0}^{-1}\right)\mathbf{v}\right\rangle \geq \langle \mathbf{v}, C_{\tau,0}^{-1}\mathbf{v}\rangle \\ &= \sum_{k=1}^K c_k^2 + \sum_{k=K+1}^N c_k^2\tau^{-2\alpha}(\sigma_{k,0}+\tau^2)^\alpha \geq \sum_{k=K+1}^N c_k^2\tau^{-2\alpha}(\sigma_{k,0}+\tau^2)^\alpha.\end{aligned}$$

Here we have used the fact that B is positive semi-definite in the first inequality. From Assumption 3.1(b), it follows that $\sigma_{k,0} \geq \theta$ for $k \geq K$, and subsequently $\sigma_{k,0} + \tau^2 \geq \theta$ for $k \geq K$. With this observation and using the expression for β in (A.3), we further continue the calculation in (A.5) to obtain the lower bound

$$\begin{aligned}\langle \mathbf{v}, (C_0^*)^{-1}\mathbf{v}\rangle &\geq \sum_{k=K+1}^N c_k^2\tau^{-2\alpha}\theta^\alpha = \tau^{-2\alpha}\theta^\alpha|(I-P_0)\mathbf{v}|^2 \\ &\geq \frac{1}{4}\tau^{-2\alpha}\theta^\alpha\left(4K/\zeta^2+1\right)^{-1}.\end{aligned}$$

Putting together the lower bounds from Cases 1 and 2 gives

$$\langle \mathbf{v}, (C_0^*)^{-1}\mathbf{v}\rangle \geq \min\left\{\gamma^{-2}(4K/\zeta^2+1)^{-1}, \frac{1}{4}\tau^{-2\alpha}\theta^\alpha(4K/\zeta^2+1)^{-1}\right\}$$

for all unit vectors $\mathbf{v}$ and constants $\gamma, \tau, \alpha > 0$. Since the trace of a matrix coincides with the sum of its eigenvalues, we conclude that

$$\operatorname{Tr}(C_0^*) \leq N\max\left\{\gamma^2(4K/\zeta^2+1), 4\tau^{2\alpha}\theta^{-\alpha}\left(4K/\zeta^2+1\right)\right\},$$

from which the desired result follows by taking $\Xi = N\left(4K/\zeta^2+1\right)\max\{1, 4\theta^{-\alpha}\}$. $\square$

A.2.2. Proof of Lemma 3.9.

Proof. Recall (3.5). Then

$$C_0^* = C_0^*\left(\frac{1}{\gamma^2}B + C_{\tau,0}^{-1}\right)C_0^* = \frac{1}{\gamma^2}C_0^*BC_0^* + C_0^*C_{\tau,0}^{-1}C^*,$$

which gives the identity

$$\mathrm{Tr}\left(\frac{1}{\gamma^2}C_0^*BC_0^*\right) = \mathrm{Tr}\left(C_0^*\right) - \mathrm{Tr}\left(C_0^*C_{\tau,0}^{-1}C_0^*\right).$$

Both C_0^* and $C_{\tau,0}^{-1}$ are positive definite and so is their product $C_0^*C_{\tau,0}^{-1}C_0^*$. Therefore, $\mathrm{Tr}\left(C_0^*C_{\tau,0}^{-1}C_0^*\right) \geq 0$ and so using Lemma 3.8 we have $\mathrm{Tr}\left(\frac{1}{\gamma^2}C_0^*BC_0^*\right) \leq \mathrm{Tr}\left(C_0^*\right) \leq \Xi \max\left\{\gamma^2, \tau^{2\alpha}\right\}$. ∎

A.2.3. Proof of Lemma 3.10.

Proof. Choose any vector $\mathbf{v} \in \mathrm{span}\{\bar{\boldsymbol{\chi}}_1, \dots, \bar{\boldsymbol{\chi}}_K\}$ and recall (3.5), the definition of C_0^*. Then

$$\begin{aligned}\left|\frac{1}{\gamma^2}C_0^*B\mathbf{v} - \mathbf{v}\right| = \left|C_0^*\left(\frac{1}{\gamma^2}B\mathbf{v} - (C_0^*)^{-1}\mathbf{v}\right)\right| &\leq \|C_0^*\|_2\left|\frac{1}{\gamma^2}B\mathbf{v} - (C_0^*)^{-1}\mathbf{v}\right| \\ &= \|C_0^*\|_2\left|C_{\tau,0}^{-1}\mathbf{v}\right| \leq \mathrm{Tr}(C_0^*)\left|C_{\tau,0}^{-1}\mathbf{v}\right|.\end{aligned}$$

Recall from Subsection 3.1 that the vectors $\bar{\boldsymbol{\chi}}_k$ are eigenvectors of L_0 corresponding to an eigenvalue of 0, and so they are also eigenvectors of $C_{\tau,0}^{-1}$ with attendant eigenvalue 1. Therefore, since $\mathbf{v} \in \mathrm{span}\left\{\bar{\boldsymbol{\chi}}_k\right\}_{k=1}^K$ it follows that $C_{\tau,0}^{-1}\mathbf{v} = \mathbf{v}$. Using this fact and Lemma 3.8 we conclude that

$$\left|\frac{1}{\gamma^2}C_0^*B\mathbf{v} - \mathbf{v}\right| \leq \Xi\max\{\gamma^2, \tau^{2\alpha}\}|\mathbf{v}|.$$

The desired bound for the vectors $\mathbf{u}_m^\dagger$ now follows trivially from Assumption 3.4. ∎

A.3. Proofs Of Lemmata 3.13, 3.14, and 3.15.

A.3.1. Proof of Lemma 3.13.

Proof. We use a similar argument to the proof of Lemma 3.8 and obtain lower bounds on the Dirichlet energy $\langle \mathbf{v}, (C_\epsilon^*)^{-1}\mathbf{v}\rangle$ for unit vectors $\mathbf{v} \in \mathbb{R}^N$. Recall $P_0 \in \mathbb{R}^{N\times N}$ denotes the projection matrix onto $\mathrm{span}\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$ and define ζ, β as in (A.3). Once again we obtain the lower bounds in two cases where $|P_0\mathbf{v}| \geq \beta$ and $|P_0\mathbf{v}| < \beta$.

The case of $|P_0\mathbf{v}| \geq \beta$ follows from identical arguments to Case 1 in the proof of Lemma 3.8. In fact, the lower bound (A.4) holds for C_ϵ^* replacing C_0^* and so whenever $|P_0\mathbf{v}| \geq \beta$ we have

$$\langle \mathbf{v}, (C_\epsilon^*)^{-1}\mathbf{v}\rangle \geq \gamma^{-2}\left(4K/\zeta^2 + 1\right)^{-1}.$$

So we focus on the case where $|P_0\mathbf{v}| < \beta$ and naturally $|(I - P_0)\mathbf{v}| \geq \sqrt{1-\beta^2}$. Let $\{(\sigma_{j,\epsilon}, \boldsymbol{\phi}_{j,\epsilon})\}_{j=1}^N$ denote the eigenpairs of L_ϵ, indexed by order of increasing eigenvalue. Note that these orthonormal eigenvectors are also eigenvectors of $C_{\tau,\epsilon}^{-1}$. We let $P_\epsilon \in \mathbb{R}^{N\times N}$ denote the projection matrix onto $\mathrm{span}\{\boldsymbol{\phi}_{1,\epsilon}, \boldsymbol{\phi}_{2,\epsilon}, \cdots, \boldsymbol{\phi}_{K,\epsilon}\}$. The key difference in this proof, compared to Case 2 in the proof of Lemma 3.8, is that we need to establish a lower bound on $|(I-P_\epsilon)\mathbf{v}|$. We show that if $\epsilon \in (0,\epsilon_0)$ for a sufficiently small constant ϵ_0, then

$$|(I-P_\epsilon)\mathbf{v}| \geq \frac{1}{2}\sqrt{1-\beta^2} = \frac{1}{2}(4K/\zeta^2+1)^{-1/2}. \tag{A.6}$$

Using (3.13) and the fact that B is positive semi-definite we can then write

$$\begin{aligned} \langle \mathbf{v}, (C^*_\epsilon)^{-1}\mathbf{v}\rangle &= \left\langle \mathbf{v}, \left(\frac{1}{\gamma^2}B + C^{-1}_{\tau,\epsilon}\right)\mathbf{v}\right\rangle \\ &\geq \langle \mathbf{v}, C^{-1}_{\tau,\epsilon}\mathbf{v}\rangle \geq \sum_{j=K+1}^{N} c^2_{j,\epsilon}\tau^{-2\alpha}(\sigma_{j,\epsilon}+\tau^2)^\alpha, \end{aligned} \tag{A.7}$$

where $c_{j,\epsilon} := \langle \mathbf{v}, \boldsymbol{\phi}_{j,\epsilon}\rangle$. By [21, Lemm A.5] the graph Laplacian L_ϵ satisfies an expansion of the form

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

where $\{\|L^{(h)}\|_2\}_{h=1}^\infty \in \ell^\infty$. Moreover, by [21, Prop. A7] and the binomial theorem we have that

$$\begin{aligned} \tau^{-2\alpha}(\sigma_{K+1,\epsilon}+\tau^2)^\alpha &\geq \tau^{-2\alpha}\left(\theta+\tau^2-\epsilon\sum_{h=1}^{\infty}\epsilon^{h-1}\|L^{(h)}\|_2\right)^\alpha \\ &> \theta^\alpha\tau^{-2\alpha}\left(1-\frac{\epsilon}{\tau^2}\sum_{h=1}^{\infty}\epsilon^{h-1}\|L^{(h)}\|_2\right)^\alpha = \theta^\alpha\tau^{-2\alpha}\left(1-\frac{\epsilon}{\tau^2}\Xi_1\right)^\alpha, \end{aligned}$$

where $\Xi_1 := \sup_{\epsilon\in(0,\epsilon_0)}\sum_{h=1}^\infty \epsilon^{h-1}\|L^{(h)}\|_2$ which is bounded provided that $\epsilon_0 < 1$. Substituting this lower bound back into (A.7) and recalling the increasing ordering of the $\sigma_{j,\epsilon}$ we obtain

$$\begin{aligned} \langle \mathbf{v}, (C^*_\epsilon)^{-1}\mathbf{v}\rangle &\geq \theta^\alpha\tau^{-2\alpha}\left(1-\frac{\epsilon}{\tau^2}\Xi_1\right)^\alpha \sum_{j=K+1}^{N} c^2_{j,\epsilon} \\ &= \theta^\alpha\tau^{-2\alpha}\left(1-\frac{\epsilon}{\tau^2}\Xi_1\right)^\alpha |(I-P_\epsilon)\mathbf{v}|^2 \geq \frac{1}{4}\theta^\alpha\tau^{-2\alpha}\left(1-\frac{\epsilon}{\tau^2}\Xi_1\right)^\alpha (4K/\zeta^2+1)^{-1}. \end{aligned}$$

Putting this bound together with the lower bound from the first case where $|P_0\mathbf{v}| \geq \beta$, we conclude that

$$\langle \mathbf{v}, (C^*_\epsilon)^{-1}\mathbf{v}\rangle \geq \min\left\{\gamma^{-2}(4K/\zeta^2+1)^{-1}, \frac{1}{4}\tau^{-2\alpha}(1-\epsilon\tau^{-2}\Xi_1)^\alpha\theta^\alpha(4K/\zeta^2+1)^{-1}\right\}$$

from which it follows that

$$\mathrm{Tr}(C^*_\epsilon) \leq N\max\left\{\gamma^2(4K/\zeta^2+1), \frac{1}{4}\tau^{2\alpha}(1-\epsilon\tau^{-2}\Xi_1)^{-\alpha}\theta^{-\alpha}(4K/\zeta^2+1)\right\}$$

provided that $\epsilon_0 > 0$ is sufficiently small which concludes the proof of the Lemma.

It remains for us to prove the bound (A.6). By [21, Prop. A.6 and proof of Prop. A.10] there exist uniform constants $\epsilon_1, \Xi_2 > 0$ so that $\forall \epsilon \in (0,\epsilon_1)$ and for any unit vector $\mathbf{v}$

$$|(I-P_\epsilon)P_0\mathbf{v}|^2 \leq \Xi_2\epsilon^2 \quad \text{and} \quad |(I-P_0)P_\epsilon\mathbf{v}|^2 \leq \Xi_2\epsilon^2,$$

implying that the range of P_ϵ and P_0 are close when ϵ is small. Therefore, using the fact that P_0 and P_ϵ are symmetric and idempotent, as well as the Cauchy-Schwarz inequality, we can write

$$\begin{aligned} |(P_0 - P_\epsilon)\,\mathbf{v}|^2 &= \langle (P_0 - P_\epsilon)\mathbf{v}, P_0\mathbf{v}\rangle - \langle (P_0 - P_\epsilon)\mathbf{v}, P_\epsilon\mathbf{v}\rangle \\ = \langle \mathbf{v}, (P_0 - P_\epsilon)\,P_0\mathbf{v}\rangle &- \langle \mathbf{v}, (P_0 - P_\epsilon)\,P_\epsilon\mathbf{v}\rangle\,\langle \mathbf{v}, (I - P_\epsilon)P_0\mathbf{v}\rangle + \langle \mathbf{v}, (I - P_0)P_\epsilon\mathbf{v}\rangle \\ &\le |\mathbf{v}|(|(I - P_\epsilon)P_0\mathbf{v}| + |(I - P_0)P_\epsilon\mathbf{v}|) \le \Xi_3\epsilon. \end{aligned}$$

The lower bound (A.6) then follows from the following calculation

$$\begin{aligned} |(I - P_\epsilon)\mathbf{v}| &= |(I - P_0)\mathbf{v} + (P_0 - P_\epsilon)\mathbf{v}| \ge \max\{0, |(I - P_0)\mathbf{v}| - |(P_0 - P_\epsilon)\mathbf{v}|\} \\ &\ge \max\left\{0, \sqrt{1-\beta^2} - (\Xi_3\epsilon)^{1/2}\right\} \ge \frac{\sqrt{1-\beta^2}}{2} = \frac{1}{2}(4K/\zeta^2 + 1)^{-1/2} \end{aligned}$$

where the last inequality holds if $\epsilon_0 \le \frac{1-\beta^2}{4\Xi_3}$. □

A.3.2. Proof of Lemma 3.15.

Proof. The proof is nearly identical to that of Lemma 3.9 and is hence omitted.□

A.3.3. Proof of Lemma 3.15.

Proof. We proceed similarly to the proof of Lemma 3.10 by choosing a vector $\mathbf{v} \in \operatorname{span}\{\bar{\boldsymbol{\chi}}_k\}_{k=1}^K$. We then have

$$\begin{aligned} \left|\frac{1}{\gamma^2}C_\epsilon^* B\mathbf{v} - \mathbf{v}\right| &= \left|C_\epsilon^*\left(\frac{1}{\gamma^2}B\mathbf{v} - (C_\epsilon^*)^{-1}\mathbf{v}\right)\right| \\ &\le \|C_\epsilon^*\|_2\left|\frac{1}{\gamma^2}B\mathbf{v} - (C_\epsilon^*)^{-1}\mathbf{v}\right| = \|C_\epsilon^*\|_2\left|C_{\tau,\epsilon}^{-1}\mathbf{v}\right|. \end{aligned}$$

Now decompose $\mathbf{v} = P_\epsilon\mathbf{v} + (I - P_\epsilon)\mathbf{v}$. Since we assumed that $\mathbf{v} \in \operatorname{span}\{\bar{\boldsymbol{\chi}}_\ell\}_{\ell=1}^K$, it follows from [21, Prop. A.6] that $|(I - P_\epsilon)\mathbf{v}| \le \Xi_3\epsilon|\mathbf{v}|$ for some $\Xi_3 > 0$ independent of ϵ, and so

$$\begin{aligned} \left|C_{\tau,\epsilon}^{-1}\mathbf{v}\right| &\le \left|C_{\tau,\epsilon}^{-1}P_\epsilon\mathbf{v}\right| + \left|C_{\tau,\epsilon}^{-1}(I - P_\epsilon)\mathbf{v}\right| \\ &\le \max_{k\le K}\frac{(\sigma_{k,\epsilon} + \tau^2)^\alpha}{\tau^{2\alpha}}|P_\epsilon\mathbf{v}| + \max_{k>K}\frac{(\sigma_{k,\epsilon} + \tau^2)^\alpha}{\tau^{2\alpha}}|(I - P_\epsilon)\mathbf{v}| \\ &\le \Xi_4\left[\left(1 + \frac{\epsilon}{\tau^2}\right)^\alpha + \epsilon\left(1 + \frac{1}{\tau^{2\alpha}}\right)\right]|\mathbf{v}|. \end{aligned}$$

The third inequality follows from the fact that the $\sigma_{k,\epsilon}$ are uniformly bounded for all $\epsilon \in (0, \epsilon_0)$ and $\epsilon_0 < 1$. In fact, by [21, Lemm. A.5], we have that

$$\begin{aligned} \sigma_{k,\epsilon} = \langle \boldsymbol{\phi}_{k,\epsilon}, L_\epsilon\boldsymbol{\phi}_{k,\epsilon}\rangle &\le |\langle \boldsymbol{\phi}_{k,\epsilon}, L_0\boldsymbol{\phi}_{k,\epsilon}\rangle| + \sum_{h=1}^\infty \epsilon^h\,|\langle \boldsymbol{\phi}_{k,\epsilon}, L_h\boldsymbol{\phi}_{k,\epsilon}\rangle| \\ &\le \|L_0\|_2 + \frac{\epsilon}{1-\epsilon}\left(\max_{h=1,2,\ldots}\|L_h\|_2\right) \le \frac{1}{1-\epsilon}\left(\max_{h=0,1,\ldots}\|L_h\|_2\right). \end{aligned}$$

Now bounding $\|C_\epsilon^*\|_2$ by $\operatorname{Tr}(C_\epsilon^*)$ and envoking Lemma 3.13 yields

$$\|C_\epsilon^*\|_2\left|C_{\tau,\epsilon}^{-1}\mathbf{v}\right| \le \Xi_0\Xi_4\max\left\{\gamma^2, \left(\frac{\tau^2}{1 - \Xi_1\epsilon/\tau^2}\right)^\alpha\right\}\left[\epsilon + \frac{\epsilon}{\tau^{2\alpha}} + \left(1 + \frac{\epsilon}{\tau^2}\right)^\alpha\right]|\mathbf{v}|.$$

The theorem follows by setting $\Xi_2 = \Xi_0\Xi_4$. □

REFERENCES

[1] S. Agapiou, S. Larsson, and A. M. Stuart, *Posterior contraction rates for the Bayesian approach to linear ill-posed inverse problems*, Stochastic Processes and their Applications, 123 (2012), https://doi.org/10.1016/j.spa.2013.05.001.

[2] M. Belkin, I. Matveeva, and P. Niyogi, *Regularization and semi-supervised learning on large graphs*, in International Conference on Computational Learning Theory, Springer, 2004, pp. 624–638.

[3] M. Belkin and P. Niyogi, *Laplacian eigenmaps and spectral techniques for embedding and clustering*, in Advances in Neural Information Processing Systems, 2002, pp. 585–591.

[4] M. Belkin, P. Niyogi, and V. Sindhwani, *Manifold regularization: A geometric framework for learning from labeled and unlabeled examples*, Journal of Machine Learning Research, 7 (2006), pp. 2399–2434.

[5] A. L. Bertozzi and A. Flenner, *Diffuse interface models on graphs for classification of high dimensional data*, SIAM Review, 58 (2016), pp. 293–328, https://doi.org/10.1137/16M1070426, https://doi.org/10.1137/16M1070426.

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

[7] L. D. Brown, M. G. Low, et al., *Asymptotic equivalence of nonparametric regression and white noise*, The Annals of Statistics, 24 (1996), pp. 2384–2398.

[8] D. Calvetti and E. Somersalo, *An introduction to Bayesian scientific computing: ten lectures on subjective computing*, vol. 2, Springer Science & Business Media, 2007.

[9] M. Dashti, K. J. Law, A. M. Stuart, and J. Voss, *MAP estimators and their consistency in Bayesian nonparametric inverse problems*, Inverse Problems, 29 (2013), p. 095017.

[10] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, arXiv preprint arXiv:1302.6989, (2013).

[11] P. Diaconis, D. Freedman, et al., *On the consistency of Bayes estimates*, The Annals of Statistics, 14 (1986), pp. 1–26.

[12] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, vol. 375, Springer Science & Business Media, 1996.

[13] D. Freedman et al., *Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters*, The Annals of Statistics, 27 (1999), pp. 1119–1141.

[14] N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs towards the Laplace–Beltrami operator*, arXiv preprint arXiv:1801.10108, (2018).

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

[16] S. Ghosal, J. K. Ghosh, A. W. Van Der Vaart, et al., *Convergence rates of posterior distributions*, Annals of Statistics, 28 (2000), pp. 500–531.

[17] E. Giné and R. Nickl, *Mathematical foundations of infinite-dimensional statistical models*, Cambridge Series In Statistical And Probabilistic Mathematics, Cambridge University Press, New York, 2016.

[18] J. Hartog and H. van Zanten, *Nonparametric bayesian label prediction on a graph*, Computational Statistics and Data Analysis, 120 (2018), pp. 111–131, https://doi.org/10.1016/j.csda.2017.11.008.

[19] J. Hartog and H. van Zanten, *Nonparametric bayesian label prediction on a large graph using truncated laplacian regularization*, Communications in Statistics - Simulation and Computation, 0 (2019), pp. 1–18, https://doi.org/10.1080/03610918.2019.1634202.

[20] F. Hoffman, B. Hosseini, A. A. Oberai, and A. M. Stuart, *Spectral analysis of weighted Laplacians arising in data clustering.* Preprint, 2019.

[21] F. Hoffman, B. Hosseini, Z. Ren, and A. M. Stuart, *Consistency of semi-supervised learning algorithms on graphs.* Preprint, 2019.

[22] J. Kaipio and E. Somersalo, *Statistical and computational inverse problems*, vol. 160, Springer Science & Business Media, 2006.

[23] A. Kirichenko and H. van Zanten, *Estimating a smooth function on a large graph by bayesian laplacian regularisation*, Electron. J. Statist., 11 (2017), pp. 891–915, https://doi.org/10.1214/17-EJS1253, https://doi.org/10.1214/17-EJS1253.

[24] G. Kostopoulos, S. Karlos, S. Kotsiantis, and O. Ragos, *Semi-supervised regression: A recent review*, Journal of Intelligent & Fuzzy Systems, (2018), pp. 1–18.

[25] Y. LeCun and C. Cortes, *MNIST handwritten digit database*, (2010), http://yann.lecun.com/exdb/mnist/.

[26] F. Lindgren, H. Rue, and J. 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 (2011), pp. 423–498.

[27] F. Monard, R. Nickl, and G. P. Paternain, *Consistent inversion of noisy non-abelian X-ray transforms*, arXiv preprint arXiv:1905.00860, (2019).

[28] A. Y. Ng, M. I. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Advances in Neural Information Processing Systems, 2002, pp. 849–856.

[29] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT press, Cambridge, 2006, http://www.gaussianprocess.org/gpml/.

[30] D. A. Spielman and S.-H. Teng, *Spectral partitioning works: Planar graphs and finite element meshes*, in Proceedings of 37th Conference on Foundations of Computer Science, IEEE, 1996, pp. 96–105.

[31] D. A. Spielman and S.-H. Teng, *Spectral partitioning works: Planar graphs and finite element meshes*, Linear Algebra and its Applications, 421 (2007), pp. 284–305.

[32] I. Steinwart, *On the influence of the kernel on the consistency of support vector machines*, Journal of Machine Learning Research, 2 (2001), pp. 67–93.

[33] I. Steinwart, *Consistency of support vector machines and other regularized kernel classifiers*, IEEE Transactions on Information Theory, 51 (2005), pp. 128–142.

[34] A. Tewari and P. L. Bartlett, *On the consistency of multiclass classification methods*, Journal of Machine Learning Research, 8 (2007), pp. 1007–1025.

[35] A. W. van der Vaart, *Asymptotic statistics*, vol. 3, Cambridge University Press, 2000.

[36] A. W. van der Vaart and J. H. van Zanten, *Rates of contraction of posterior distributions based on Gaussian process priors*, The Annals of Statistics, 36 (2008), pp. 1435–1463.

[37] V. Vapnik, *Statistical learning theory*, Wiley, 1998.

[38] U. Von Luxburg, *A tutorial on spectral clustering*, Statistics and Computing, 17 (2007), pp. 395–416.

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

[40] M. Wu and B. Schölkopf, *Transductive classification via local learning regularization*, in Artificial Intelligence and Statistics, 2007, pp. 628–635.

[41] Q. Wu and D.-X. Zhou, *Analysis of support vector machine classification*, Journal of Computational Analysis & Applications, 8 (2006).

[42] L. Zelnik-Manor and P. Perona, *Self-tuning spectral clustering*, in Advances in Neural Information Processing Systems, 2005, pp. 1601–1608.

[43] X. Zhu, *Semi-supervised Learning with Graphs*, PhD thesis, Pittsburgh, PA, USA, 2005. AAI3179046.

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

[45] X. Zhu, J. Lafferty, and Z. Ghahramani, *Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions*, in ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining, 2003, pp. 58–65.

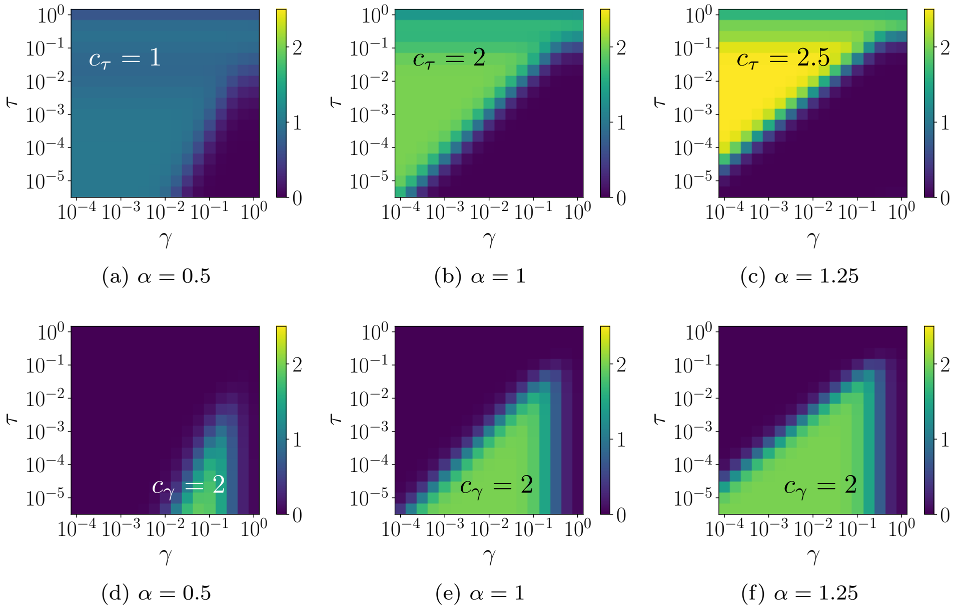
[46] X. J. Zhu, *Semi-supervised learning literature survey*, Tech. Report TR1530, University of Wisconsin-Madison, Computer Sciences Department, 2005.

Appendix B. Supplemental Material - Numerical Demonstration of Lemmata.

B.1. Numerics In Support Of Lemmata 3.8 to 3.10. In Figures 4 and 5 we present numerics that illustrate the convergence results for Lemmata 3.8 and 3.10 respectively. These lemmata respectively bound the first and third terms of the decomposition of $\mathcal{I}$:

$$\mathcal{I}(\gamma, \alpha, \tau) = M\mathrm{Tr}(C_0^*) + \frac{M}{\gamma^2}\mathrm{Tr}(C_0^* B C_0^*) + \sum_{m=1}^{M} \left| \frac{1}{\gamma^2} C_0^* B \mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger \right|^2 .$$

Numerics for the middle term, $1/\gamma^2\mathrm{Tr}(C_0^* B C_0^*)$, are omitted since the corresponding bound in Lemma 3.9 is derived from the bound found for $\mathrm{Tr}(C_0^*)$ in Lemma 3.8, and exhibit nearly identical behavior numerically. The top panels in Figure 4 show



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 1.25$

(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 1.25$

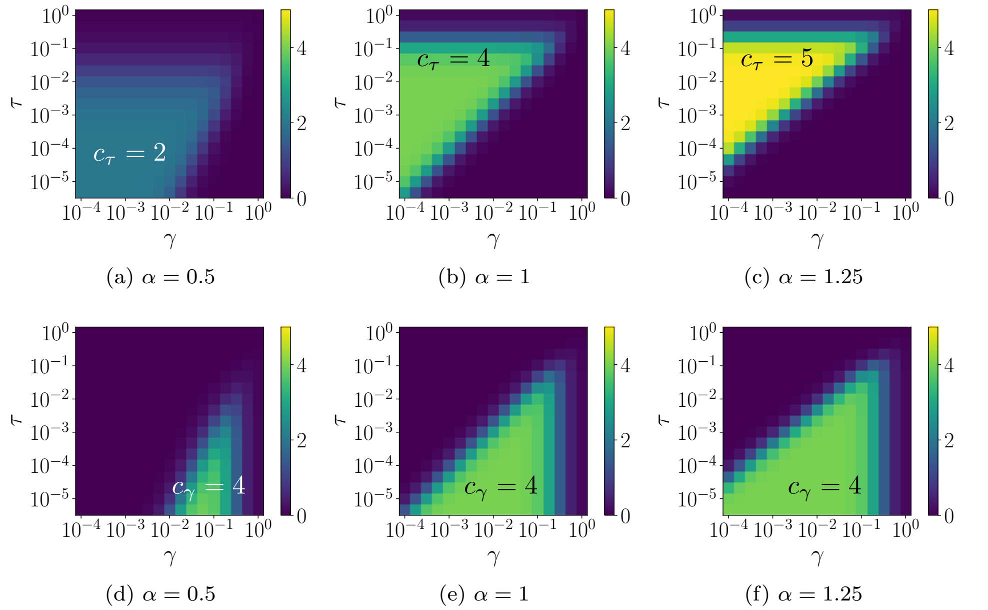
Fig. 4. *A numerical demonstration of Lemma 3.8 on the synthetic dataset with $\epsilon = 0$. The top panels showcase the numerical estimates of the logarithmic slope $c_\tau := \log(\mathrm{Tr}(C_0^*))/\partial \log(\tau)$ for different α values and the bottom panels showcase the numerical estimates of the logarithmic slope $c_\gamma := \partial \log(\mathrm{Tr}(C_0^*))/\partial \log(\gamma)$. In the dark blue region $c_\tau, c_\gamma \approx 0$, indicating that $\mathrm{Tr}(C_0^*)$ stays flat with respect to the respective variable τ or γ; the slope of the brighter regions is denoted on each figure. The transition between the dark and bright regions occurs approximately at $\tau = \gamma^{1/\alpha}$.*

the estimated rate of convergence of $\mathrm{Tr}(C_0^*)$ in terms of τ in the log-log scale, while the bottom panels show the estimated rate of convergence in terms of γ in the log-log scale. Figure 5 likewise shows the estimated rate of convergence $\left|\frac{1}{\gamma^2} C_0^* B \mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger\right|^2$ in the parameters τ and γ. From Figure 4, we read that in the region where $\gamma^2 \ll \tau^{2\alpha}$, $\partial \log(\mathrm{Tr}(C_0^*))/\partial \log(\tau)$ stays close to 2α whereas $\partial \log(\mathrm{Tr}(C_0^*))/\partial \log(\gamma)$ is approximately 0. In the region where $\tau^{2\alpha} \ll \gamma^2$, we observe that $\partial \log(\mathrm{Tr}(C_0^*))/\partial \log(\tau)$ is close to 0 whereas $\partial \log(\mathrm{Tr}(C_0^*))/\partial \log(\gamma)$ is around 2. These results confirm our bound in Lemma 3.8.

In Figure 5, we read that in the region where $\gamma^2 \ll \tau^{2\alpha}$, $\partial \log(|C_0^* B \mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)$ $/\partial \log(\tau)$ stays close to 4α whereas $\partial \log(|C_0^* B \mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial \log(\gamma)$ is approximately 0. In the region where $\tau^{2\alpha} \ll \gamma^2$, we observe that $\partial \log(|C_0^* B \mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)$ $/\partial \log(\tau)$ is close to 0 whereas $\partial \log(|C_0^* B \mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial \log(\gamma)$ is around 4. These results confirm our bounds presented in Lemma 3.10.

B.2. Numerics In Support Of Lemmata 3.13 to 3.15. In Figures 6 and 7 we present numerics that illustrate the convergence results for Lemmata 3.13 and 3.15 respectively. These lemmata respectively bound the first and third terms of the decomposition of $\mathcal{I}$:

$$\mathcal{I}(\gamma, \alpha, \tau, \epsilon) = M \mathrm{Tr}(C_\epsilon^*) + \frac{M}{\gamma^2} \mathrm{Tr}(C_\epsilon^* B C_\epsilon^*) + \sum_{m=1}^{M} \left| \frac{1}{\gamma^2} C_\epsilon^* B \mathbf{u}_m^\dagger - \mathbf{u}_m^\dagger \right|^2 .$$



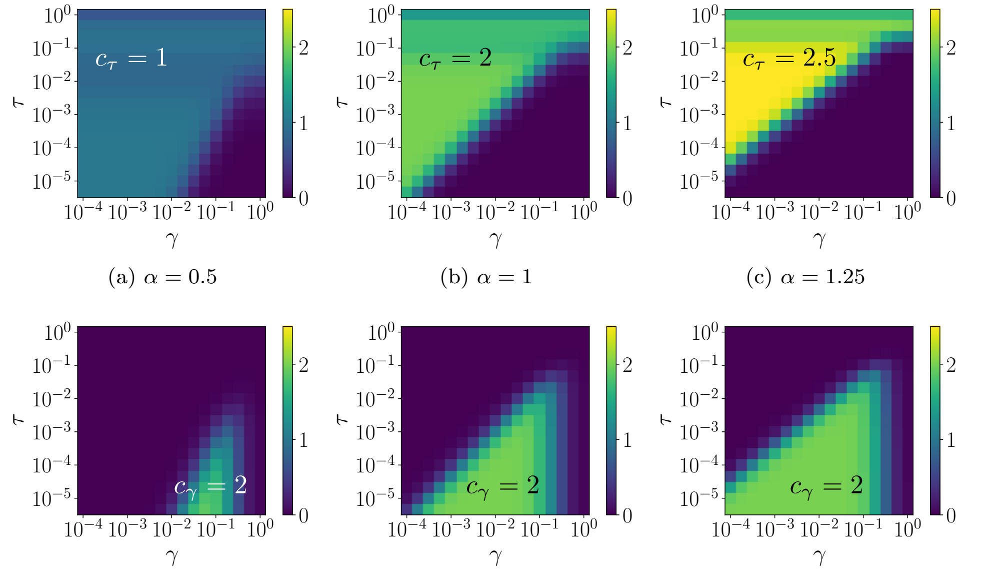
(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 1.25$

(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 1.25$

FIG. 5. *A numerical demonstration of Lemma 3.10 on the synthetic dataset with $\epsilon = 0$. The top panels showcase the numerical estimates of the logarithmic slope $c_\tau := \log(|C_0^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial\log(\tau)$ for different α values and the bottom panels showcase the numerical estimates of the logarithmic slope $c_\gamma := \partial\log(|C_0^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial\log(\gamma)$. In the dark blue region $c_\tau, c_\gamma \approx 0$, indicating that $|C_0^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2$ stays flat with respect to the respective variable τ or γ; the slope of the brighter regions is denoted on each figure. The transition between the dark and bright regions occurs approximately at $\tau = \gamma^{1/\alpha}$.*

Again, we omit numerics for the middle term in this decomposition since the corresponding bound in Lemma 3.14 is derived from the bound found for $\mathrm{Tr}(C_\epsilon^*)$ in Lemma 3.13 and exhibit nearly identical behavior numerically. Just as in Figures 2 and 3, we have set the scaling $\epsilon = \tau^{\max\{2,2\alpha\}}$. The top panels in Figure 6 show the estimated rate of convergence of $\mathrm{Tr}(C_\epsilon^*)$ in terms of τ in the log-log scale, while the bottom panels show the estimated rate of convergence in terms of γ in the log-log scale. Figure 7 likewise shows the estimated rate of convergence $|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|$ in the parameters ϵ and γ. From Figure 6, we read that in the region where $\gamma^2 \ll \tau^{2\alpha}$, $\partial\log(\mathrm{Tr}(C_\epsilon^*))/\partial\log(\tau)$ stays close to 2α whereas $\partial\log(\mathrm{Tr}(C_\epsilon^*))/\partial\log(\gamma)$ is approximately 0. In the region where $\tau^{2\alpha} \ll \gamma^2$, we observe that $\partial\log(\mathrm{Tr}(C_\epsilon^*))/\partial\log(\tau)$ is close to 0 whereas $\partial\log(\mathrm{Tr}(C_\epsilon^*))/\partial\log(\gamma)$ is around 2. These results confirm our bound presented in Lemma 3.13.

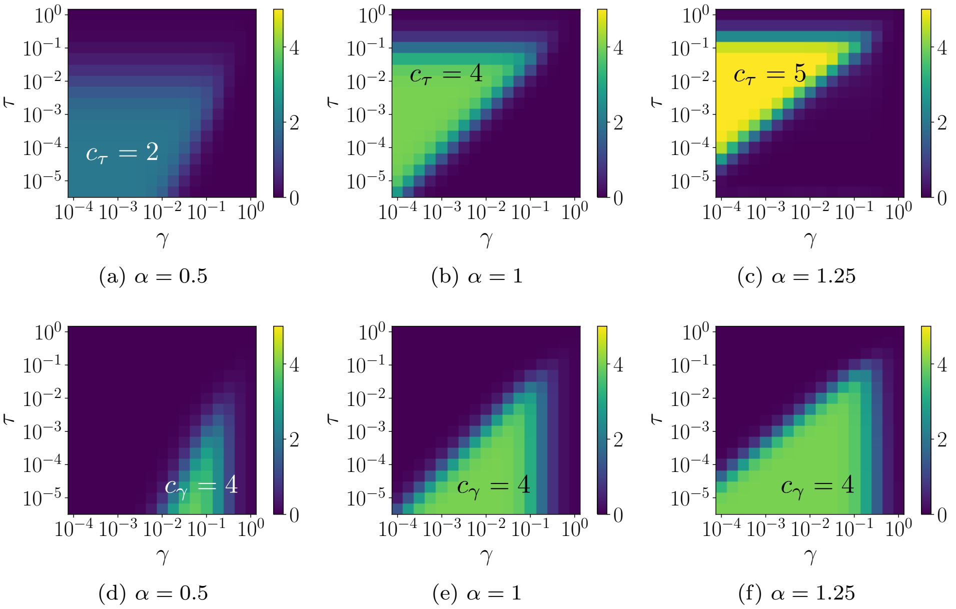
In Figure 7, we read that in the region where $\gamma^2 \ll \tau^{2\alpha}$, $\partial\log(|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)$ $/\partial\log(\tau)$ stays close to 4α whereas $\partial\log(|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial\log(\gamma)$ is approximately 0. In the region where $\tau^{2\alpha} \ll \gamma^2$, we observe that $\partial\log(|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)$ $/\partial\log(\tau)$ is close to 0 whereas $\partial\log(|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial\log(\gamma)$ is around 4. These results confirm our bounds presented in Lemma 3.15.



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 1.25$

(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 1.25$

Fig. 6. *A numerical demonstration of Lemma 3.13 on the synthetic dataset with $\epsilon = \tau^{2\alpha}$. The top panels showcase the numerical estimates of the logarithmic slope $c_\tau := \log(\mathrm{Tr}(C^*_\epsilon))/\partial \log(\tau)$ for different α values and the bottom panels showcase the numerical estimates of the logarithmic slope $c_\gamma := \partial \log(\mathrm{Tr}(C^*_\epsilon))/\partial \log(\gamma)$. In the dark blue region $c_\tau, c_\gamma \approx 0$, indicating that $\mathrm{Tr}(C^*_\epsilon)$ stays flat with respect to the respective variable τ or γ; the slope of the brighter regions is denoted on each figure. The transition between the dark and bright regions occurs approximately at $\tau = \gamma^{1/\alpha}$.*



(a) $\alpha = 0.5$ (b) $\alpha = 1$ (c) $\alpha = 1.25$

(d) $\alpha = 0.5$ (e) $\alpha = 1$ (f) $\alpha = 1.25$

FIG. 7. *A numerical demonstration of Lemma 3.15 on a synthetic dataset with* $\epsilon = \tau^{\max\{2,2\alpha\}}$*. The top panels showcase the numerical estimates of the logarithmic slope* $c_\tau := \log(|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial\log(\tau)$ *for different* α *values and the bottom panels showcase the numerical estimates of the logarithmic slope* $c_\gamma := \partial\log(|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2)/\partial\log(\gamma)$*. In the dark blue region* $c_\tau, c_\gamma \approx 0$*, indicating that* $|C_\epsilon^* B\mathbf{u}_m^\dagger/\gamma^2 - \mathbf{u}_m^\dagger|^2$ *stays flat with respect to the respective variable* τ *or* γ*; the slope of the brighter regions is denoted on each figure. The transition between the dark and bright regions occurs approximately at* $\tau = \gamma^{1/\alpha}$*.*