A Sampling Theory Perspective of Graph-based Semi-supervised Learning

Aamir Anis, Student Member, IEEE, Aly El Gamal, Member, IEEE, Salman Avestimehr, Senior Member, IEEE, and Antonio Ortega, Fellow, IEEE

Abstract—Graph-based methods have been quite successful in solving unsupervised and semi-supervised learning problems, as they provide a means to capture the underlying geometry of the dataset. It is often desirable for the constructed graph to satisfy two properties: first, data points that are similar in the feature space should be strongly connected on the graph, and second, the class label information should vary smoothly with respect to the graph, where smoothness is measured using the spectral properties of the graph Laplacian matrix. Recent works have justified some of these smoothness conditions by showing that they are strongly linked to the semi-supervised smoothness assumption and its variants. In this work, we reinforce this connection by viewing the problem from a graph sampling theoretic perspective, where class indicator functions are treated as bandlimited graph signals (in the eigenvector basis of the graph Laplacian) and label prediction as a bandlimited reconstruction problem. Our approach involves analyzing the bandwidth of class indicator signals generated from statistical data models with separable and nonseparable classes. These models are quite general and mimic the nature of most real-world datasets. Our results show that in the asymptotic limit, the bandwidth of any class indicator is also closely related to the geometry of the dataset. This allows one to theoretically justify the assumption of bandlimitedness of class indicator signals, thereby providing a sampling theoretic interpretation of graph-based semi-supervised classification.

I. INTRODUCTION

The abundance of unlabeled data in various machine learning applications, along with the prohibitive cost of labeling, has led to growing interest in semi-supervised learning. This paradigm deals with the task of classifying data points in the presence of very little labeling information by relying on the geometry of the dataset. Assuming that the features are well-chosen, a natural assumption in this setting is to consider the marginal density \( p(x) \) of the feature vectors to be informative about the labeling function \( f(x) \) defined on the points. This assumption is fundamental to the semi-supervised learning problem both in the classification and the regression settings, and is also known as the semi-supervised smoothness assumption \([1]\), which states that the label function is smoother in regions of high data density. There also exist other similar variants of this assumption specialized for the classification setting, namely, the cluster assumption \([2]\) (points in a cluster are likely to have the same class label) or the low density separation assumption \([3]\) (decision boundaries pass through regions of low data density). Most present day algorithms for semi-supervised learning rely on one or more of these assumptions to predict the unknown labels.

In practice, graph-based methods have been found to be quite suitable for geometry-based learning tasks, primarily because they provide an easy way of exploiting information from the geometry of the dataset. These methods involve constructing a distance-based similarity graph whose vertices represent the data points and whose edge weights are in general a decreasing function of distance between them. The key assumption here is that the label function is “smooth” over the graph, in the sense that labels of vertices do not vary much over edges with high weights (i.e., edges that connect close or similar points). There are numerous ways of quantitatively imposing smoothness constraints over label functions defined on vertices of a similarity graph. Most graph-based semi-supervised classification algorithms incorporate one of these criteria as a penalty against the fitting error in a regularization problem, or as a constraint term while minimizing the fitting error in an optimization problem. For example, a commonly used measure of smoothness for a label function \( f \) is the graph Laplacian regularizer \( f^T L f \) (\( L \) being the graph Laplacian), and many algorithms involve minimizing this quadratic energy function while ensuring that \( f \) satisfies the known set of labels \([2]. [4]\). There also exist higher-order variants of this measure known as iterated graph Laplacian regularizers \( f^T L^m f \), that have been shown to make the problem more well-behaved \([5]\). On the other hand, a spectral theory based classification algorithm restricts \( f \) to be spanned by the first few eigenvectors of the graph Laplacian \([6]. [7]\), that are known to form a representation basis for smooth functions on the graph. In each of the examples, the criterion enforces smoothness of the labels over the graph – a lower value of the regularizer \( f^T L f \), and a smaller number of leading eigenvectors to model \( f \) imply that vertices that are close neighbors on the graph are more likely to have the same label.

A more recent approach, derived from the emerging field of Graph Signal Processing (GSP) \([8]\), considers the semi-supervised learning problem from the perspective of sampling theory for graph signals \([9]. [12]\). It involves treating the class label function \( f \) as a bandlimited graph signal, and
label prediction as a bandlimited reconstruction problem. The advantage of this approach is that one can also analyze, using sampling theory, the label complexity of semi-supervised classification, that is, the minimum fraction of labeled examples required for perfect prediction of the unknown labels. A key ingredient in this formulation is the bandwidth $\omega(f)$ of signals on the graph, which is defined as the largest Laplacian eigenvalue for which the projection of the signal over the corresponding eigenvector is non-zero. Signals with lower bandwidth tend to be smoother on the graph, and are useful for modeling label functions over similarity graphs. Label prediction using bandlimited reconstruction then involves estimating a label function that minimizes prediction error on the known set under a bandwidth constraint. This can also be carried out without explicitly computing the eigenvectors of the Laplacian, and has been shown to be quite competitive in comparison to state-of-the-art graph-based semi-supervised learning methods [13].

Although graph-based semi-supervised learning methods are well-motivated, their connection to the underlying geometry of the dataset is not clearly understood so far in a theoretical sense. Recent works focused on justifying these approaches by exploring their geometrical interpretation in the limit of infinitely available unlabeled data. This is typically done by assuming a probabilistic generative model for the dataset and analyzing the graph smoothness criteria in the asymptotic setting for certain commonly-used graph construction schemes. For example, it has been shown that for data points drawn from a smooth distribution with an associated smooth label function (i.e., the regression setting), the graph Laplacian regularizer converges in the limit of infinite data points to a density-weighted variational energy functional that penalizes large variations of the labels in high density regions [5], [7], [14]–[17]. A similar connection ensues for semi-supervised learning problems in the classification setting (i.e., when labels are discrete in the feature space). If points drawn from a smooth distribution are separated by a smooth boundary into two classes, then the graph cut for the partition converges to a weighted volume of the boundary [3], [18]. This is consistent with the low density separation assumption – a low value of the graph cut implies that the boundary passes through regions of low data density.

However, these results cannot be applied for providing an interpretation of the sampling theoretic approach to learning, since we lack knowledge about the convergence behavior of the bandwidth of class indicator signals. An interpretation of this approach would help complete our theoretical understanding of graph-based semi-supervised learning approaches and strengthen their link with the semi-supervised smoothness assumption and its variants. Therefore, in this work, we seek answers for the following questions:

- What is the connection between the bandwidth of class indicator signals over the similarity graph and the underlying geometry of the data set?
- What is the interpretation of the bandlimited reconstruction approach for label prediction?
- How many labeled examples does one require for perfect prediction?

To answer these questions, our work analyzes the asymptotic behavior of an iterated Laplacian-based bandwidth estimator for class indicator signals on similarity graphs constructed from a statistical model for the feature vectors. To make our analysis as general as possible, we consider two data models: separable and nonseparable. These generative models are quite practical and can be used to mimic most datasets in the real world. The separable model assumes that data points are independently drawn from an underlying probability distribution in the feature space and each class is separated from the others by a smooth boundary. On the other hand, the nonseparable model assumes a mixture distribution for the data where the data points are drawn independently with certain probability from separate class conditional distributions. We also introduce a notion of “boundaries” for classes in the nonseparable model in the form of overlap regions (i.e., the region of ambiguity), defined as the set of points where the probability of belonging and not belonging to a class are both non-zero. This definition is quite practical and useful for characterizing the geometry of such datasets.

Using the data points, we consider a specific graph construction scheme that applies the Gaussian kernel over Euclidean distances between feature vectors for computing their similarities (our analysis can be generalized easily to arbitrary kernels under simple assumptions). In order to compute the bandwidth of any signal on the graph, we define an estimator based on the iterated Laplacian regularizer. A significant portion of this paper focuses on analyzing the stochastic convergence of this bandwidth estimate (using variance-bias decomposition) in the limit of infinite data points for any class indicator signal on the graph. The analysis in our work suggests a novel sampling theoretic interpretation of graph-based semi-supervised learning and the main contributions can be summarized as follows:

- **Relationship between bandwidth and data geometry.** For the separable model, we show that under certain rate conditions, the bandwidth estimate for any class converges to the supremum of the data density over the class boundary. Similarly, for the nonseparable model, we show that the bandwidth estimate converges to the supremum of the density over the overlap region.
- **Interpretation of bandlimited reconstruction.** Using the geometrical interpretation of the bandwidth, we conclude that bandlimited reconstruction allows one to choose the complexity of the hypothesis space while predicting unknown labels (i.e., a larger bandwidth allows more complex class boundaries).
- **Quantification of label complexity.** We also show that the bandwidth of class indicator signals is closely linked theoretically to the fraction of labeled points required for perfect classification which is in turn related to the geometry of the data.

Our analysis has significant implications: Firstly, class indicator signals have a low bandwidth if class boundaries lie in regions of low data densities, that is, the semi-supervised assumption holds for graph-based methods. And secondly, our analysis also helps quantify the impact of bandwidth and data geometry in semi-supervised learning problems. Specifically, it
enables us to theoretically assert that for the sampling theoretic approach to graph-based semi-supervised learning, the label complexity (minimum fraction of labeled points required) of learning classifiers matches the theoretical estimate and is indeed lower if the boundary lies in regions of low data density, as demonstrated empirically in earlier works [6, 7].

The rest of this paper is organized as follows: In Section II, we formally introduce the statistical data models and the graph construction scheme for analysis. In Section III, we review prior work and underline their connections with our work. In Section IV we state our main results and outline their implications. In Section V we prove the major building blocks for our results and leave the details of smaller components to the Appendix. We finally conclude with numerical validation in Section VI, followed by discussion and an outline of future work. It is worth noting that the bandwidth convergence result for the separable model and an interpretation of bandlimited reconstruction were given in our preliminary work [19]. This paper presents complete formal proofs for those results, extends them to the nonseparable model, and also considers the label complexity problem.

II. Preliminaries

A. Data models

1) The separable model: In this model, we assume that the dataset consists of a pool of \( n \) random, \( d \)-dimensional feature vectors \( \mathcal{X} = \{X_1, X_2, \ldots, X_n\} \) drawn independently from some probability density function \( p(x) \) supported on \( \mathbb{R}^d \) (this is assumed for simplicity, the analysis can be extended to arbitrary manifolds \( \mathcal{M} \subset \mathbb{R}^d \), but would more technically involved). To simplify our analysis, we also assume that \( p(x) \) is bounded from above, Lipschitz continuous and twice differentiable. We assume that a smooth hypersurface \( \partial S \), with radius of curvature lower bounded by a constant \( \tau \), splits \( \mathbb{R}^d \) into two disjoint classes \( S \) and \( S' \), with indicator functions \( 1_S(x) : \mathbb{R}^d \to \{0, 1\} \) and \( 1_{S'}(x) : \mathbb{R}^d \to \{0, 1\} \). This is illustrated in Figure 1a. Thus, the \( n \)-dimensional class indicator signal for class \( S \) is given by \( 1_S \in \{0, 1\}^n \) such that \( 1_S(i) = 1_S(X_i) \), i.e., the \( i \)-th entry of \( 1_S \) is 1 if \( X_i \in S \) and 0 otherwise.

2) The nonseparable model: In this model, we assume that each class has its own conditional distribution supported on \( \mathbb{R}^d \) (that may or may not overlap with other distributions of other classes). The data set consists of a pool of \( n \) random and independent \( d \)-dimensional feature vectors \( \mathcal{X} = \{X_1, X_2, \ldots, X_n\} \) drawn independently from any of the distributions \( p_i(x) \) with probabilities \( \alpha_i \), such that \( \sum_i \alpha_i = 1 \).

For our analysis, we consider a class denoted by an index \( A \) with selection probability \( \alpha_A \), class conditional distribution \( p_A(x) \) and an \( n \)-dimensional indicator vector \( 1_A \) whose \( i \)-th component takes value 1 if \( X_i \) is drawn from class \( A \). This model is illustrated in Figure 1b. Further, we denote by \( \alpha_{A^c} = 1 - \alpha_A \) the probability that a point does not belong to \( A \) and by \( p_{A^c}(x) = \sum_i \alpha_i p_i(x)/\alpha_A \) the density of all such points. The marginal distribution of data points is then given by the mixture density

\[
p(x) = \alpha_A p_A(x) + \alpha_{A^c} p_{A^c}(x).
\]

Once again, to simplify our analysis, we assume that all distributions are Lipschitz continuous, bounded from above and twice differentiable in \( \mathbb{R}^d \). Next, we introduce the notion of a “boundary” for classes in the nonseparable model as follows: for class \( A \), we define its overlap region \( \partial A \) as

\[
\partial A = \{x \in \mathbb{R}^d \mid p_A(x)p_{A^c}(x) > 0\}
\]

Intuitively, \( \partial A \) can be considered as the region of ambiguity, where both points belonging and not belonging to \( A \) co-exist. In other words, \( \partial A \) can be thought of as a “boundary” that separates the region where points can only belong to \( A \) from the region where points can never belong to \( A \). Since class indicator signals on graphs will change values only within the overlap region, one would expect that the indicators will be smoother if there are fewer data points within this region. We shall show later that this is indeed the case, both theoretically and experimentally.

B. Graph construction

Using the \( n \) feature vectors, we construct an undirected distance-based similarity graph where nodes represent the data points and edge weights are proportional to their similarity, given by the Gaussian kernel:

\[
w_{ij} = K_{\sigma^2}(X_i, X_j) = \frac{1}{(2\pi\sigma^2)^{d/2}}e^{-\|X_i-X_j\|^2/2\sigma^2},
\]

where \( \sigma \) is the variance (bandwidth) of the Gaussian kernel. Further, we assume \( w_{ii} = 0 \), i.e., the graph does not have self-loops. The adjacency matrix of the graph \( W \) is an \( n \times n \) symmetric matrix with elements \( w_{ij} \), while the degree matrix is a diagonal matrix with elements \( D_{ii} = \sum_j w_{ij} \). We define the graph Laplacian as \( L = \tfrac{1}{2}(D - W) \). Normalization by \( n \) ensures that the norm of \( L \) is stochastically bounded as \( n \) grows. Since the graph is undirected, \( L \) is a symmetric matrix with non-negative eigenvalues \( 0 \leq \lambda_1 \leq \cdots \leq \lambda_n \) and an orthogonal set of corresponding eigenvectors \( \{u_1, \ldots, u_n\} \). It is known that for a larger eigenvalue \( \lambda \), the corresponding eigenvector \( u \) exhibits greater variation when plotted over the nodes of the graph [8].

C. Bandwidth of graph signals

The bandwidth \( \omega(f) \) of any signal \( f \) on the graph is defined as the largest eigenvalue for which the projection of the signal on the corresponding eigenvector is non-zero, i.e.,

\[
\omega(f) = \max_i \{\lambda_i \mid u_i^T f > 0\}.
\]

Ideally, finding the bandwidth \( \omega(f) \) of a graph signal \( f \) requires computing the eigenvectors \( \{u_i\} \) and the corresponding projections \( f_i = u_i^T f \). However, analyzing the convergence
of these coefficients is technically challenging. Therefore, we resort to the following estimate of the bandwidth \([12]\):

\[
\omega_m(f) = \left( \frac{f^T L_m^f f}{f^T f} \right)^{1/m},
\]

where we call \(\omega_m(f)\) the \(m^{th}\)-order bandwidth estimate. It can be shown that the bandwidth estimates satisfy the property: for all \(0 < m_1 < m_2\), \(\omega_{m_1}(f) \leq \omega_{m_2}(f) \leq \omega(f)\). In other words, \(\{\omega_m(f)\}\) forms a monotonically improving sequence of estimates of the true bandwidth \(\omega(f)\). Thus, we have the following relation:

\[
\forall f, \ \omega(f) = \lim_{m \to \infty} \omega_m(f).
\]

Analyzing the convergence of \(\omega_m(1_S)\) and \(\omega_m(1_A)\) as \(n \to \infty\), \(\sigma \to 0\) and \(m \to \infty\) constitutes the main subject for the rest of this paper. Specifically, we relate these quantities to the underlying data distribution \(p(x)\) and class boundaries (the hypersurface \(\partial S\) in the separable case and the overlap region \(\partial A\) in the nonseparable case).

Note that the limit in (6) holds in a point-wise sense. This means that analyzing the convergence of the bandwidth estimates \(\omega_m(1_S)\) and \(\omega_m(1_A)\) as \(n \to \infty\) gives only an idea about the convergence of actual bandwidths \(\omega(1_S)\) and \(\omega(1_A)\) as \(n \to \infty\). Specifically, it does not imply convergence of \(\omega(1_S)\) and \(\omega(1_A)\) to the same values as \(\omega_m(1_S)\) and \(\omega_m(1_A)\), since the limits are not interchangeable unless (6) holds in a uniform sense. However, based on our experiments and results on label complexity, we believe that our intuition is accurate, i.e., the convergence results hold for the actual bandwidths, not only their estimates. We leave the analysis of this intricacy for future work.

### D. Bandlimited interpolation for classification

Bandwidth plays an important role in the sampling theoretic approach for semi-supervised learning. In this approach, one finds a label assignment by minimizing the error over the known set, while ensuring that the resulting class indicator vector is bandlimited over the similarity graph, i.e.,

\[
\min \ ||f(L) - y(L)||^2 \ \text{subject to} \ \omega(f) < \omega_L \quad (7)
\]

where \(L\) denotes the set of known labels, \(y\) denotes the true class labels and \(f(L)\) and \(y(L)\) denote the values of \(f\) and \(y\) on the set \(L\) respectively. The constraint restricts the hypothesis space by constraining it to a set of bandlimited signals with bandwidth less than \(\omega_L\), which is equivalent to enforcing smoothness of the labels over the graph. One can also use sampling theory to set \(\omega_L\) as the cutoff frequency associated with the sampling set \([12]\), as considered in \([13], [20]\).

Note that the bandwidth-based approach for semi-supervised learning extends the Fourier eigenvector approach suggested in \([6], [7]\) by allowing the estimation of the complexity of the bandlimited space via \(\omega_L\). Further, the label prediction can be carried out without explicitly computing the eigenvectors of \(L\). It is based on iterative and alternate projections onto convex sets and can be implemented in an efficient manner via graph filtering operations \([21]\).

### III. Related work and connections

Existing graph-based semi-supervised learning and spectral clustering methods have been justified by analyzing the convergence of graph-based smoothness measures (such as the graph cut and the Laplacian regularizer) for various graph construction schemes in two different settings — classification and regression. The classification setting assumes that labels indicate class memberships and are discrete, typically with \(1/0\) values. Note that both the separable and nonseparable data models considered in our paper are in the classification setting. On the other hand, in the regression setting, one allows the class label signal \(f\) to be smooth and continuous with soft values, i.e., \(f \in \mathbb{R}^n\) and later applies some thresholding mechanism to infer class memberships. For example, in the two class problem, one can assign +1 and −1 to the two classes and threshold \(f\) at 0. Convergence analysis of smoothness measures in this setting requires different scaling conditions than the classification setting, and leads to fundamentally different limiting quantities that require differentiability of the label functions. A summary of convergence results in the literature for both settings is presented in Table 1. Although these results do not focus on analyzing the bandwidth of class indicator signals, the proof techniques used in this paper are inspired by some of these works. We review them in this section and discuss their connections to our work.

#### A. Classification setting

Prior work under this setting assumes the separable data model where the feature space is partitioned by smooth decision boundaries into different classes. When \(m = 1\), the bandwidth estimate \(\omega_m(1_S)\) for the separable model in our
TABLE I: Related convergence results in the literature under different data models and graph construction schemes. All models assume that the distributions are smooth (at least twice-differentiable). Further, the graph Laplacian is defined as $L = \frac{1}{n}(D-W)$ in all cases. [18] also studies convergence of graph cuts for weighted $k$-nearest neighbor and $r$-neighborhood graphs which we do not include for brevity.

| Work | Data model | Graph model | Quantity | Convergence regime | Limit (within constant scaling factor) |
|------|------------|-------------|----------|-------------------|---------------------------------------|
| Narayanan et al [5] | $p(x)$ supported on manifold $\mathcal{M} \subset \mathbb{R}^d$, separated into $S$ and $S^c$ by smooth hypersurface $\partial S$ | Normalized Gaussian weights $w_{ij} = \frac{\sigma}{\sqrt{d_i d_j}}$ | $\frac{1}{n} \sigma \int_{S} L_1 \sigma$ | $n \to \infty$, $\sigma \to 0$ | $\int_{\partial S} p(s) ds$ |
| Maier et al [18] | $p(x)$ supported on $\mathcal{M} \subset \mathbb{R}^d$, separated into $S$ and $S^c$ by hyperplane $\partial S$ | $r$-neighborhood, unweighted | $\frac{1}{n} \int_{\partial S} L_1 \sigma$ | $n \to \infty$, $r \to 0$ | $\int_{\partial S} p^2(s) ds$ |
| Bousquet et al [14], Hein [15] | $p(x)$ and $f(x)$ supported on $\mathbb{R}^d$ | Fully-connected, Gaussian weights $\frac{1}{n} \int_{\partial S} L_1 \sigma$ | $n \to \infty$, $\sigma \to 0$ | $\int_{\partial S} p^2(s) ds$ |
| Zhou et al [5] | Uniformly distributed on $d$-dim. submanifold $\mathcal{M}$ | Fully-connected, Gaussian weights $\frac{1}{n} \int_{\partial S} L_1 \sigma$ | $n \to \infty$, $\sigma \to 0$ | $\int_{\partial S} p^2(s) ds$ |
| El Alaoui et al [22] | $p(x)$ supported on $[0,1]^d$ | Fully-connected, Gaussian weights $\frac{1}{n}\int_{\partial S} L_1 \sigma$ | $n \to \infty$, $\sigma \to 0$ | $\int_{\partial S} p^2(s) ds$ |
| Current work | $p(x)$ supported on $\mathbb{R}^d$, separated into $S$ and $S^c$ by smooth hypersurface $\partial S$ | Fully-connected, Gaussian weights $\frac{1}{n}\int_{\partial S} L_1 \sigma$ | $n \to \infty$, $\sigma \to 0$ | $\int_{\partial S} p^2(s) ds$ |
| Drawn from $p_A(x)$ and $p_{A^c}(x)$ supported on $\mathbb{R}^d$ with probabilities $\alpha_A$ and $\alpha_{A^c}$ | Fully-connected, Gaussian weights $\frac{1}{n}\int_{\partial S} L_1 \sigma$ | $n \to \infty$, $\sigma \to 0$ | $\int_{\partial S} p^2(s) ds$ |

A similar result for a similarity graph constructed with normalized weights $w_{ij} = \frac{\sigma}{\sqrt{d_i d_j}}$ was shown earlier for an arbitrary hypersurface $\partial S$ in [3], where $d_i$ denotes the degree of node $i$. In this case, normalization of the graph weights results in convergence to $\frac{1}{\sqrt{2\pi}} \int_{\partial S} p(s) ds$. The results in [3], [18] aim to provide an interpretation for spectral clustering – up to some scaling, the empirical cut value converges to a weighted volume of the boundary. Thus, spectral clustering is a means of performing low density separation on a finite sample drawn from a distribution in feature space.

Although these works provide inspiration for the proof techniques used for analyzing the separable model in this paper, they cannot be directly used in the convergence analysis of $\omega_m(1_S)$ for $m > 1$, which is the main focus of our paper. Additionally, our work is the first to propose and analyze the nonseparable model in the classification setting, i.e., convergence results for $\omega_m(1_A)$.

**B. Regression setting**

To predict the labels of unknown samples in the regression setting, one generally minimizes the graph Laplacian regularizer $f^T L f$ subject to the known label constraints [4]:

$$\min_{f} f^T L f \text{ such that } f(L) = y(L),$$

One particular convergence result in this setting assumes that $n$ data points are drawn i.i.d. from $p(x)$ and are labeled by sampling a smooth function $f(x)$ on $\mathbb{R}^d$. Here, the graph...
Laplacian regularizer $f^T L f$ can be shown to converge in the asymptotic limit under the conditions $\sigma \to 0$ and $n\sigma^d \to \infty$ as in \cite{14, 15}:

$$\frac{1}{n\sigma^2} f^T L f \xrightarrow{p} C \int_{\mathbb{R}^d} \|\nabla f(x)\|^2 p^2(x) dx,$$  \hspace{1cm} (11)

where for each $n$, $f$ is the $n$-dimensional label vector representing the values of $f(x)$ at the $n$ sample points, $\nabla$ is the gradient operator and $C$ is a constant factor independent of $n$ and $\sigma$. The right hand side of the result above is a weighted Dirichlet energy functional that penalizes variation in the label function weighted by the data distribution. Similar to the justification of spectral clustering, this result justifies the formulation in \cite{10} for semi-supervised classification: given label constraints, the predicted label function must vary little in regions of high density. The work of \cite{13, 23} also generalizes the result for arbitrary kernel functions used in defining graph weights, and data distributions defined over arbitrary manifolds in $\mathbb{R}^d$. Similar convergence results have also been derived for the higher-order Laplacian regularizer $f^T L^m f$ obtained from uniformly distributed data \cite{5}. In this case, it was shown that for data points obtained from a uniform distribution on a $d$-dimensional submanifold $\mathcal{M} \subset \mathbb{R}^N$ such that $\text{Vol}(\mathcal{M}) = 1$ and $2m$-differentiable functions $f(x)$, one has as $n \to \infty$, $\sigma \to 0$,

$$\frac{1}{n \sigma^m} f^T L^m f \xrightarrow{p} C \int_{\mathcal{M}} f(x) \Delta^m f(x) dx,$$  \hspace{1cm} (12)

where $\Delta$ is the Laplace operator and $\sigma_n = n^{-1/(2d+4+\alpha)}$ is a vanishing sequence with $\alpha > 0$. Extensions for non-uniform probability distributions $p(x)$ over the manifold can be obtained using the weighted Laplace-Beltrami operator \cite{16, 17}. More recently, an $\ell_p$-based Laplacian regularization has been proposed for imposing smoothness constraints in semi-supervised learning problems \cite{22}. This is similar to higher-order regularization but is defined as $J_p(f) = \sum_{i,j \in E} w_{ij}^p |f_i - f_j|^p$, where $w_{ij} = K(|X_i - X_j|/\sigma)$ and $K(\cdot)$ is a smoothly decaying Kernel function. It has been shown for a bounded density $p(x)$ defined on $[0,1]^d$ that for every $p \geq 2$, as $n \to \infty$, $\sigma \to 0$,

$$\frac{1}{n^2 \sigma^{d+p}} J_p(f) \xrightarrow{p} C \int_{[0,1]^d} \|\nabla f(x)\|^2 p^2(x) dx.$$  \hspace{1cm} (13)

Although our work also uses higher powers of $L$ in the expressions for $\omega_m(1_S)$ and $\omega_m(1_A)$, we cannot use the convergence results in \cite{12, 13} since they are only applicable for smooth functions (i.e., differentiable up to certain order) on $\mathbb{R}^d$. Specifically, these results cannot be applied for the bandwidth of discrete 0/1 class indicator functions.

To summarize, the results in the literature mostly pertain to convergence analysis of variants of the graph cut or the graph Laplacian regularizer for different models of data and graph construction schemes, and do not provide insight into the convergence of bandwidths of discrete 0/1 class indicator signals. In contrast, we analyze bandwidth expressions involving these class indicator signals and higher powers of $L$, and for the first time, extend it to a nonseparable data model. As opposed to other smoothness measures considered earlier, analyzing the bandwidth allows us to interpret graph-based semi-supervised learning using the sampling theorem \cite{12} and provide a quantitative insight into label complexity based on data geometry.

IV. MAIN RESULTS AND DISCUSSION

A. Interpretation of bandwidth and bandlimited reconstruction

We first show that under certain conditions, the bandwidth estimates of class indicator signals, over the distance-based similarity graph described earlier, converge to quantities that are functions of the underlying distribution and the class boundary for both data models. This convergence is achieved under the following asymptotic regime:

1) Increasing size of dataset: $n \to \infty$.
2) Shrinking neighborhood volume: $\sigma \to 0$.
3) Improving bandwidth estimates: $m \to \infty$.

Note that an increasing size of the dataset (Condition 1) is required for the stochastic convergence of the bandwidth estimate. Condition 2 ensures that the limiting values are concise and have a simple interpretation in terms of the data geometry. Intuitively, Condition 2 ensures that as the number of data points increases, one looks at a smaller neighborhood around each data point, as a result, the degree of each node in the graph does not blow up. Finally, Condition 3 leads to improving values of the bandwidth estimate. The convergence results are precisely stated in the theorems below:

Theorem 1. If $n \to \infty$, $\sigma \to 0$ and $m \to \infty$ while satisfying the following rate conditions

1) $(n\sigma^{md})/(mC^m) \to \infty$, where $C = 2/(2\pi)^{d/2}$,
2) $m/n \sigma \to 0$,
3) $m\sigma^2 \to 0$,
4) $\sigma^{1/m} \to 1$,

then for the separable model, one has

$$\omega_m(1_S) \xrightarrow{p} \sup_{s \in \partial S} p(s),$$  \hspace{1cm} (14)

where “$p$” denotes convergence in probability.

Theorem 2. If $n \to \infty$, $\sigma \to 0$ and $m \to \infty$ while satisfying the following rate conditions

1) $(n\sigma^{md})/(mC^m) \to \infty$, where $C = 2/(2\pi)^{d/2}$,
2) $m/n \to 0$,
3) $m\sigma^2 \to 0$,

then for the non-separable model, one has

$$\omega_m(1_A) \xrightarrow{p} \sup_{x \in \partial A} p(x).$$  \hspace{1cm} (15)

The dependence of the results on the rate conditions will be explained later in the proofs section. An example of parameter choices that allow all the scaling laws to hold simultaneously is illustrated in the following corollary:

Corollary 1. Equations (14) and (15) hold if for each value of $n$, we choose $m$ and $\sigma$ as follows:

$$m = [n_0 \log n]^p,$$  \hspace{1cm} (16)

$$\sigma = \sigma_0 n^{-d/m},$$  \hspace{1cm} (17)
for constants $m_0, \sigma_0 > 0$, $1/2 < y < 1$ and $0 < x < 1$. [. .] indicates taking the nearest integer value.

Theorems 1 and 2 give an explicit connection between bandwidth estimates of class indicator signals and the underlying geometry of the dataset. This interpretation forms the basis of justifying the choice of bandwidth as a smoothness constraint in graph-based learning algorithms. Theorem 1 suggests that for the separable model, if the boundary $\partial S$ passes through regions of low probability density, then the bandwidth $\omega(1_S)$ is low. A similar conclusion is suggested for the nonseparable model from Theorem 2, i.e., if the density of data points in the overlap region $\partial A$ is low, then the bandwidth $\omega(1_A)$ is low. In other words, low density of data in the boundary regions leads to smooth indicator functions.

From our results, we also get an intuition behind the smoothness constraint imposed in the bandlimited reconstruction approach (7) for semi-supervised learning. Basically, enforcing smoothness on classes in terms of indicator bandwidth ensures that the learning algorithm chooses a boundary passing through regions of low data density in the separable case. Similarly, in the nonseparable case, it ensures that variations in labels occur in regions of low density. Further, the bandwidth cutoff $\omega_L$ effectively imposes a constraint on the complexity of the hypothesis space – a larger value increases the size of the hypothesis space and results in choosing more complex boundaries.

As a special case of our analysis, we also get a convergence result for the graph cut in the nonseparable model analogous to the results of (18) for the separable model. Note that the cut in this case equals the sum of weights of edges connecting points that belong to class $A$ to points that do not belong to class $A$, i.e.,

$$\text{Cut}(A, A^c) = \sum_{x_i \in A, x_j \in A^c} w_{ij} = 1^t_A L 1_A.$$ (18)

With this definition, we have the following result:

**Theorem 3.** If $n \to \infty$, $\sigma \to 0$ and $n \sigma^{d+1} \to \infty$, then

$$\frac{1}{n}\text{Cut}(A, A^c) \to p \int \alpha_A \alpha_{A^c} p_A(x)p_{A^c}(x) dx.$$ (19)

The result above indicates that if the overlap between the conditional distributions of a particular class and its complement is low, then the value of the graph cut is lower. This justifies the use of spectral clustering in the context of nonseparable models.

### B. Label complexity of SSL

In the context of semi-supervised learning, we define the label complexity as the minimum fraction of labeled examples required for perfectly predicting the labels of the unlabeled data points. This quantity essentially is an indicator of how “good” the semi-supervised problem is, i.e., how much help do we get from geometry while learning. A low label complexity is indicative of a favorable situation, where one is able to learn from only a few known labels by exploiting data geometry. In the following discussion, we first estimate the theoretical label complexities of the data models we consider, and then show that the expected label complexity of the sampling theoretic approach to learning exactly matches these values in the asymptotic limit.

**Theoretical label complexities:** A simple way to compute the label complexity, for the data models we consider, is to find the fraction of points belonging to a region that fully encompasses the boundary. To formalize this, let us define the following two regions in $\mathbb{R}^d$:

$$\mathcal{X}_S = \{x : p(x) \leq \sup_{s \in \partial S} p(s)\},$$ (20)

$$\mathcal{X}_A = \{x : p(x) \leq \sup_{x \in \partial A} p(x)\}$$ (21)

Note that by definition, $\partial S$ is fully contained in $\mathcal{X}_S$ and $\partial A$ is fully contained in $\mathcal{X}_A$ (see Figure 2 for an example in $\mathbb{R}^3$). To perfectly reconstruct $\mathcal{L}_S$ and $\mathcal{L}_A$, it is sufficient to know the labels of all points in $\mathcal{X}_S$ and $\mathcal{X}_A$ respectively, as this strategy removes all ambiguity in labeling the two classes. Based on this, we arrive at the following conclusions:

**Observation 1.** The theoretical label complexity of learning $\mathcal{L}_S$ and $\mathcal{L}_A$ in the asymptotic limit are $P(\mathcal{X}_S)$ and $P(\mathcal{X}_A)$ respectively, where $P(E) = \int_E p(x) dx$.

**Label complexity of graph-based learning:** Using our results, we can show that the same label complexities hold for the graph-based sampling theoretic approach to semi-supervised classification. In this context, label complexity can be seen as the fraction of samples required for perfectly reconstructing a signal on the similarity graph. It is known that the fraction of samples required for perfectly reconstructing a bandlimited signal cannot be more than the fraction of eigenvalues of the Laplacian below the signal’s bandwidth (12). Since our bandwidth convergence results relate the bandwidth of indicators for the two data models with data geometry, we only need to asymptotically relate the number of eigenvalues of $L$ below any constant in terms of data geometry. This is achieved through the following result:

**Theorem 4.** Let $N_L(t)$ be the number of eigenvalues of $L$ below a constant $t$. Then, as $n \to \infty$ and $\sigma \to 0$, we have

$$\mathbb{E}\left\{\frac{1}{n}N_L(t)\right\} \to P\left(\{x : p(x) \leq t\}\right),$$ (22)

**Proof.** See Section V-D.

Substituting the bandwidth convergence results from Theorems 1 and 2 (i.e., $t = \omega_m(1_S)$ and $t = \omega_m(1_A)$), we immediately get the desired value of the expected label complexity of graph-based semi-supervised learning:

**Theorem 5.** If the conditions in Theorems 1 and 2 hold, then the expected label complexities of bandlimited reconstruction for the separable and nonseparable models are given as

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E}\{N_L(\omega_m(1_S))\} \to P(\mathcal{X}_S),$$ (23)

$$\lim_{n \to \infty} \frac{1}{n} \mathbb{E}\{N_L(\omega_m(1_A))\} \to P(\mathcal{X}_A).$$ (24)

The following remarks are in order:
1) Note that Theorem 4 and Theorem 5 can be strengthened by proving convergence of \( \frac{1}{n}N_E(t) \) rather than its expected value. This requires further analysis, which we leave for future work. The result in Theorem 5 also encourages us to conjecture that the convergence results for bandwidth estimates also hold for the convergence of the bandwidth itself.

2) This result further strengthens the connection between graph-based learning methods and the semi-supervised smoothness assumption, since one can conclude that the number of labeled examples required for perfect prediction depends on the geometry of the data around the boundary. A low value of the density at the boundary results in a lower label complexity.

3) One might ask what is the advantage of using graph-based methods for semi-supervised learning, if we can predict the class labels by the simple labeling strategy: predict the class labels by the simple labeling strategy. Based on our experiments, we observe that the number of labeled examples required for perfect prediction depends on the geometry of the data around the boundary. The power of graph-based methods and the semi-supervised learning methods for semi-supervised learning, if we can consider the random variable: 

\[
V = \frac{1}{n}1_R^T L^m 1_R,
\]

in order to apply a concentration inequality for \( V \), we first re-write it in the form of a U-statistic by regrouping terms in the summation in order to remove repeated indices, as given in [20]: 

\[
V = \frac{1}{n^{m+1}} \sum_{i_1, i_2, \ldots, i_{m+1}} g^* (X_{i_1}, X_{i_2}, \ldots, X_{i_{m+1}}),
\]

The above expansion has the form of a V-statistic. Details on how to explicitly write the summation are given in Appendix A. Note that \( g \) is composed of a sum of \( 2^m \) terms; each a product of \( m \) kernel functions that are non-negative. Therefore, 

\[
g \leq 2^m ||K||_m^m = \left( \frac{2}{(2\pi)^{d/2}} \right)^m = C_m \sigma^{md}.
\]

In order to apply a concentration inequality for \( V \), we first re-write it in the form of a U-statistic by regrouping terms in the summation in order to remove repeated indices, as given in [20]: 

\[
V = \frac{1}{n^{m+1}} \sum_{(n,m+1)} g^* (X_{i_1}, X_{i_2}, \ldots, X_{i_{m+1}}),
\]

A partial sketch of the proof for the separable model is also provided in our parallel work [19]; here we provide the complete proof.
where \( \sum_{(n,m+1)} \) denotes summation over all ordered \((m+1)\)-tuples of distinct indices taken from the set \{1, \ldots, n\}, \( n_{(m+1)} = n(n-1) \ldots (n-m) \) is the number of \((m+1)\)-permutations of \( n \) and \( g^* \) is a convex combination of specific instances of \( g \) that absorbs repeating indices (see supplementary material for a complete expansion):

\[
g^*(x_1, x_2, \ldots, x_{m+1}) = \frac{n_{(m+1)}}{p_{m+1}} g(x_1, x_2, \ldots, x_{m+1})
\]

+ (terms with repeated indices). (31)

Therefore, \( g^* \) has the same upper bound as that of \( g \) derived in (29). Moreover, using the fact that \( \mathbb{E}\{V\} = \mathbb{E}\{g^*\} \), we can bound the variance of \( g^* \) as

\[
\text{Var}(g^*) \leq \|g^*\|_{\infty} \mathbb{E}\{g^*\} = \frac{C m}{\sigma^m} \mathbb{E}\{V\}. \quad (32)
\]

Finally, plugging in the bound and variance of \( g^* \) in Bernstein’s inequality for \( U\)-statistics as stated in [23], [24], we arrive at the desired result of (27). ■

Note that as \( n \rightarrow 0 \) and \( \sigma \rightarrow 0 \) with rates satisfying \((n\sigma^m)/(m\sigma^m) \rightarrow \infty \), we have \( P(|V - \mathbb{E}\{V\}| > \epsilon) \rightarrow 0 \) for all \( \epsilon > 0 \). The continuous mapping theorem then allows us to conclude that \( V^{1/m} \overset{p}{\rightarrow} (\mathbb{E}\{V\})^{1/m} \).

B. Convergence of the bias term for the separable model

To evaluate the convergence of bias terms, we shall require the following properties of the \( d \)-dimensional Gaussian kernel:

**Lemma 2.** If \( p(x) \) is twice differentiable, then

\[
\int K_{\sigma^2}(x,y)p(y)dy = p(x) + O(\sigma^2). \quad (33)
\]

**Proof.** Using the substitution \( y = x + t \) followed by a Taylor series expansion about \( x \), we have

\[
\int K_{\sigma^2}(x,y)p(y)dy = \int \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\|t\|^2/2\sigma^2} p(x + t) dt
\]

\[
= \int \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\|t\|^2/2\sigma^2} \left( p(x) + t^T \nabla p(x) + \frac{1}{2} t^T \nabla^2 p(x) t + \ldots \right) dt
\]

\[
= p(x) + O(\sigma^2),
\]

where \( \text{Tr}(\cdot) \) denotes the trace of a matrix, and the third step follows from simple component-wise integration. ■

**Lemma 3.** If \( p(x) \) is twice differentiable, then

\[
\int K_{\sigma^2}(x,z)K_{\sigma^2}(z,y)p(z)dz
\]

\[
= K_{(\alpha + b)\sigma^2}(x,y) \left( p \left( \frac{bx + a y}{a + b} \right) + O(\sigma^2) \right). \quad (34)
\]

**Proof.** Note that

\[
K_{(\alpha + b)\sigma^2}(x,z)K_{(\alpha + b)\sigma^2}(z,y)
\]

\[
= \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\|x-z\|^2/2\sigma^2} \left( \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\|y-z\|^2/2\sigma^2} \right)
\]

\[
= \frac{1}{(2\pi(a+b)^2)^{d/2}} e^{-\|x-y\|^2/2(a+b)^2} \left( \frac{1}{(2\pi a^2)^{d/2}} \right) e^{-\|z\|^2/2a^2}
\]

\[
= K_{(\alpha+b)\sigma^2}(x,y) K_{(\alpha+b)\sigma^2} \left( \frac{bx + ay}{a + b}, z \right).
\]

Therefore, we have

\[
\int K_{(\alpha+b)\sigma^2}(x,z)K_{(\alpha+b)\sigma^2}(z,y)p(z)dz
\]

\[
= K_{(\alpha+b)\sigma^2}(x,y) \int K_{(\alpha+b)\sigma^2} \left( \frac{bx + ay}{a + b}, z \right) p(z)dz
\]

\[
= K_{(\alpha+b)\sigma^2}(x,y) \left( p \left( \frac{bx + ay}{a + b} \right) + O(\sigma^2) \right),
\]

where the last step follows from Lemma 2. ■

In order to prove convergence for the separable model, we need the following results:

**Lemma 4.** If \( p(x) \) is Lipschitz continuous, then for a smooth hypersurface \( \partial S \) that divides \( \mathbb{R}^d \) into \( S_1 \) and \( S_2 \), and whose radius has curvature that is bounded by \( \tau \geq 0 \),

\[
\lim_{\sigma \rightarrow 0} \frac{1}{\sigma} \int_{S_1} \int_{S_2} K_{\sigma^2}(x_1, x_2) p^\alpha(x_1)p^\beta(x_2) dx_1 dx_2
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_{\partial S} p^\alpha(s) ds, \quad (35)
\]

where \( \alpha \) and \( \beta \) are positive integers. Moreover, for positive integers \( a, b \), and \( \alpha, \beta, \alpha', \beta' \) such that \( \alpha + \beta = \alpha' + \beta' = \gamma \), we have:

\[
\lim_{\sigma \rightarrow 0} \frac{1}{\sigma} \int_{S_1} \int_{S_1} \left[ K_{\sigma^2}(x_1, x_2) p^\alpha(x_1)p^\beta(x_2) - K_{\sigma^2}(x_1, x_2) p^\alpha(x_1)p^\beta(x_2) \right] dx_1 dx_2
\]

\[
= \frac{\sqrt{\beta} - \sqrt{a}}{\sqrt{2\pi}} \int_{\partial S} p^{\gamma}(s) ds. \quad (36)
\]

**Proof.** See Appendix B ■

We now prove the deterministic convergence of \( \mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m 1_S \right\} \) in the following lemma:

**Lemma 5.** As \( n \rightarrow \infty, \sigma \rightarrow 0 \) such that \( m/n \rightarrow 0, \sigma^2 \rightarrow 0 \), we have

\[
\frac{1}{\sigma} \mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m 1_S \right\} \rightarrow \frac{t(m)}{\sqrt{2\pi}} \int_{\partial S} p^{m+1}(s) ds, \quad (37)
\]

where \( t(m) = \sum_{r=1}^{m} \binom{m-1}{r-1} (-1)^r (\sqrt{r} + 1 - \sqrt{r}) \).

**Proof.** We evaluate \( \mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m 1_S \right\} \) term by term by expanding \( \mathbf{L}^m \) as \( (\mathbf{D} - \mathbf{W})^m \). Details on the intermediate
steps of this expansion are given in Appendix A. Using \(\sum n\), we have for the first two terms of the expansion:

\[
\frac{1}{\sigma} \mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_S \right\} = \frac{1}{\sigma} \int_S \int_S K_{\sigma^2}(x,y) p^m(x)p(y) \, dx \, dy + O(\alpha) + O\left( \frac{m}{n\sigma} \right). \tag{38}
\]

For the rest of the terms, we also require the use of \(\sum n\). However, in this case, we encounter several terms of the form \(p(\theta x + (1 - \theta)y)\) for some \(\theta \in [0, 1]\). Since \(m/s^2 \to 0\) and \(p(x)\) is assumed to be Lipschitz continuous, we can approximate such terms by \(p(x)\) or \(p(y)\). Therefore, for all terms in the expansion of \((\mathbf{D} - \mathbf{W})^{m-1}\) containing \(r > 1\) occurrences of \(\mathbf{W}\) (there are \(\binom{m-1}{r}\) such terms), repeated use of \(\sum n\), \(\sum n\) gives:

\[
\frac{1}{\sigma} \mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^{m-1-r} \mathbf{W}^r \mathbf{(D - W)} \mathbf{1}_S \right\} = \frac{1}{\sigma} \int_S \int_S K_{\sigma^2}(x,y) p^m(x)p^\beta(y) \, dx \, dy - \int_S \int_S K_{(r+1)\sigma^2}(x,y) p^{\alpha}(x)p^{\beta}(y) \, dx \, dy + O(\alpha) + O\left( \frac{m}{n\sigma} \right), \tag{39}
\]

where \(\alpha + \beta = \alpha' + \beta' = m + 1\) and \([D^{m-1-r}, W^r]\) denotes an expression containing \(r\) occurrences of \(W\) and \(m-1-r\) occurrences of \(D\). Now, using Lemma 2, we conclude that the right hand sides of \(\sum n\) and \(\sum n\) converge to \(\frac{1}{2\pi} \int_{S^1} p^{m+1}(s) \, ds\) and \(\frac{\sqrt{r+r-\sqrt{r^2}}}{\sqrt{2\pi}} \int_{S^1} p^{m+1}(s) \, ds\), respectively, as \(s \to 0\) and \(m/n \to 0\). Putting everything together in the expansion of \(\mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_S \right\}\), we get the desired result.

Since \(s^{1/m} \to 1\), we have

\[
\left( \mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_S \right\} \right)^{1/m} = s^{1/m} \left( \frac{1}{\sigma} \mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_S \right\} \right)^{1/m} \to \frac{t(m)}{\sqrt{2\pi}} \int_{S^1} p^{m+1}(s) \, ds \right)^{1/m}. \tag{40}
\]

Finally, we note that as \(m \to \infty\), we have

\[
\left( \frac{t(m)}{\sqrt{2\pi}} \int_{S^1} p^{m+1}(s) \, ds \right)^{1/m} \to \sup_{s \in S^1} p(s). \tag{41}
\]

Therefore, we conclude for the separable model

\[
\omega_m(1) \to \sup_{s \in S^1} p(s). \tag{42}
\]

C. Convergence of bias term for the nonseparable model

For the nonseparable model, we need to prove convergence of \(\mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_A \right\}\). This is illustrated in the following lemma:

**Lemma 6.** As \(n \to \infty\), \(\sigma \to 0\) such that \(m/n \to 0\), \(m\sigma^2 \to 0\), we have

\[
\mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_A \right\} \to \int \alpha_A \alpha_{A'} p_A(x)p_{A'}(x) p^{m-1}(x) \, dx. \tag{43}
\]

**Proof.** Similar to the proof of Lemma 5 we evaluate \(\mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_A \right\}\) term by term by expanding \(\mathbf{L}^m\) as \((\mathbf{D} - \mathbf{W})^{m-1}\). Using \(\sum n\), \(\sum n\), \(\sum n\), repeatedly, we have for the first two terms of the expansion:

\[
\mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_A \right\} = \int \alpha_A \alpha_{A'} p_A(x)p_{A'}(x) p^{m-1}(x) \, dx + O(\sigma^2) + O\left( \frac{m}{n} \right). \tag{44}
\]

Further, for all terms in the expansion of \((\mathbf{D} - \mathbf{W})^{m-1}\) containing \(r > 1\) occurrences of \(\mathbf{W}\) (there are \(\binom{m-1}{r}\) such terms), repeated use of \(\sum n\), \(\sum n\) gives:

\[
\mathbb{E}\left\{ \frac{1}{n} \mathbf{L}^m \mathbf{1}_A \right\} \to \int \alpha_A \alpha_{A'} p_A(x)p_{A'}(x) p^{m-1}(x) \, dx + O(\sigma^2) + O\left( \frac{m}{n} \right). \tag{45}
\]

Therefore, as \(\sigma \to 0\), \(m/n \to 0\), we get the desired result. □

We finally note that as \(m \to \infty\), we have

\[
\left( \frac{\int \alpha_A \alpha_{A'} p_A(x)p_{A'}(x) p^{m-1}(x) \, dx}{\int A p(x) \, dx} \right)^{1/m} \to \sup_{x \in \partial A} p(x). \tag{46}
\]

Therefore, we conclude for the nonseparable model

\[
\omega_m(1) \to \sup_{x \in \partial A} p(x). \tag{47}
\]

Note that Lemma 6 for the special case of \(m = 1\) yields

\[
\frac{1}{n} \mathbf{L}^m \mathbf{1}_A \to \int \alpha_A \alpha_{A'} p_A(x)p_{A'}(x) \, dx, \tag{48}
\]

which proves Theorem 4.

D. Proof of Theorem 4

We begin by recalling the definition of the empirical spectral distribution (ESD) of \(L\): \(\mu_n(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - \lambda_i)\), where \(\{\lambda_i\}\) are the eigenvalues of \(L\). For each \(x\), \(\mu_n(x)\) is a function of \(X_1, \ldots, X_n\), and thus a random variable. Note that the fraction of eigenvalues of \(L\) below a constant \(t\), and its expected value can be computed from the ESD as

\[
\frac{1}{n} \mathbb{N}_L(t) = \int_0^t \mu_n(x) \, dx, \tag{49}
\]

\[
\mathbb{E}\left\{ \frac{1}{n} \mathbb{N}_L(t) \right\} \to \int_0^t \mathbb{E}\{\mu_n(x)\} \, dx. \tag{50}
\]

Therefore, to understand the behavior of the expected fraction of eigenvalues of \(L\) below \(t\), we need to analyze the convergence of the expected ESD in the asymptotic limit. The idea is to show the convergence of the moments of \(\mathbb{E}\{\mu_n(x)\}\) to the moments of a limiting distribution \(\mu(x)\). Then, by a standard convergence result, \(\mathbb{E}\{\mu_n(I)\} \to \mu(I)\) for intervals \(I\). More precisely, let the \(\Rightarrow\) symbol denote weak convergence
of measures, then we use the following result that follows from
the Weierstrass approximation theorem:

Lemma 7. Let \( \mu_n \) be a sequence of probability measures
and \( \mu \) be a compactly supported probability measure. If
\( \int x^m \mu_n(dx) \to \int x^m \mu(dx) \) for all \( m \geq 1 \), then \( \mu_n \Rightarrow \mu \).

We then use the following result on equivalence of different
notions of weak convergence of measures [25] Theorem
25.2] in order to prove our result for cumulative distribution
functions.

Lemma 8. \( \mu_n \Rightarrow \mu \) if and only if \( \mu_n(A) \to \mu(A) \) for every
\( \mu \)-continuity set \( A \).

We begin by writing the \( m \)th moment of \( E \{ \mu_n(x) \} \):

\[
\int x^m \mathbb{E} \{ \mu_n(x) \} \, dx = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \{ \lambda^m \} = \mathbb{E} \left\{ \frac{1}{n} \text{Tr}(L^m) \right\}.
\]

Now, note that \( L^m = (D - W)^m = D^m + \sum_{k=1}^{m} \binom{m}{k} D^{m-k} W^k \), where \( D^{m-k}, W^k \) denotes product
terms with \( m-k \) occurrences of \( D \) and \( k \) occurrences of \( W \).
Therefore, we have for the right hand side of (51):

\[
\mathbb{E} \left\{ \frac{1}{n} \text{Tr}(L^m) \right\} = \int \left( \int K(x_{i_1}, x_{i_2}) p(x_{i_2}) dx_{i_2} \right) \ldots \int K(x_{i_1}, x_{i_{m+1}}) p(x_{i_{m+1}}) dx_{i_{m+1}} \int p(x_i) dx_i \right) \text{Tr}(L^m) + O \left( \frac{m}{n} \right) \text{(expected value of other terms)}.
\]

Using (33) repeatedly in the equation above, we get:

\[
\int x^m \mathbb{E} \{ \mu_n(x) \} \, dx = \int p^{m+1}(x) \, dx + O \left( \frac{m}{n} \right) + O (\sigma^2).
\]

Therefore, as \( n \to \infty \) and \( \sigma \to 0 \), we have:

\[
\int x^m \mathbb{E} \{ \mu_n(x) \} \, dx = \int p^m(x) p(x) \, dx.
\]

From the right hand side of the equation above, we conclude
that the \( m \)th moment of the expected ESD of \( L \) converges
to the \( m \)th moment of the distribution of a random variable
\( Y = p(X) \), where \( p(x) \) is the probability density function
of \( X \). Moreover, since \( p_Y(y) \) has compact support, \( E \{ \mu_n(x) \} \)
converges weakly to the probability density function of \( p_Y(y) \).
Hence, the following can be said about the expected fraction
of eigenvalues of \( L \):

\[
\mathbb{E} \left\{ \frac{1}{n} N_L(t) \right\} = \int_0^t \mathbb{E} \{ \mu_n(x) \} \, dx.
\]

This proves our claim in Theorem 4. Note that, to prove the
stochastic convergence of the fraction itself rather than its
expected value, we would need a condition similar to those
in Theorems 1 and 2 to hold for each moment. In that case, \( \sigma \)
will go to 0 in a prohibitively slow fashion. We believe that this
is an artifact of the methods we employ for proving the result.

Hence, our conjecture is that the convergence result holds for
\( \frac{1}{n} N_L(t) \) itself, and we leave the analysis of this statement for
future work.

VI. NUMERICAL VALIDATION

We now present simple numerical analyses to validate
our results and demonstrate their usefulness in practice. For
simulating the separable model, we first consider a data
distribution based on a 2D Gaussian Mixture Model (GMM)
with two Gaussians: \( \mu_1 = [-1 \ 0], \Sigma_1 = 0.25I \) and
\( \mu_2 = [1 \ 0], \Sigma_2 = 0.16I \), and mixing proportions \( \alpha_1 = 0.4 \)
and \( \alpha_2 = 0.6 \) respectively. The probability density function
is illustrated in Figure 5. Next, we evaluate the claim of
Theorem 1 on five boundaries, described in Table II. These
boundaries are depicted in Figure 4 and are illustrative of
typical separation assumptions such as linear or non-linear
and low or high density.

For simulating the nonseparable model, we first construct
the following smooth (twice-differentiable) 2D probability
density function

\[
q(x, y) = \begin{cases} \frac{2}{\pi} \left[ 1 - (x^2 + y^2)^2 \right], & x^2 + y^2 \leq 1 \\ 0, & x^2 + y^2 > 1 \end{cases}
\]

Note that datapoints \((X, Y)\) can be sampled from this
distribution by setting the coordinates \( X = \sqrt{1 - U^{1/2} \cos(2\pi V)}, \)
\( Y = \sqrt{1 - U^{1/2} \sin(2\pi V)} \), where \( U, V \sim \text{Uniform}(0, 1) \).
We then use \( q(x, y) \) to define a nonseparable 2D model
with mixture density \( p(x, y) = \alpha_A p_A(x, y) + \alpha_A^c p_{A^c}(x, y) \),
where \( p_A(x, y) = q(x - 0.75y, y), p_{A^c}(x, y) = q(x + 0.75y, y) \)
and \( \alpha_A = \alpha_{A^c} = 0.5 \). The probability density function
is illustrated in Figure 3. The overlap region or boundary \( \partial A \)
for this model is given by

\[
\partial A = \{(x, y) : (x - 0.75)^2 + y^2 < 1 \}
\]

Further, for this model, we have \( \sup_{\partial A} p(x) = 0.2517 \).

In our first experiment, we validate the statements of
Theorems 1 and 2 by comparing the left and right hand sides
of (14) and (15) for corresponding boundaries. This is carried
out in the following way: we draw \( n = 2500 \) points from each
model and construct the corresponding similarity graphs using
\( \sigma = 0.1 \). Then, for the boundaries \( \partial S_1 \) in the separable
model and \( \partial A \) in the nonseparable model, we carry out the following
steps:

| Boundary | Description | \( \sup_{\partial S_2} p(s) \) |
|----------|-------------|-----------------|
| \( \partial S_1 \) | \( x = 0 \) | 0.0607 |
| \( \partial S_2 \) | \( x = -1 \) | 0.2547 |
| \( \partial S_3 \) | \( x = y^2 - 1 \) | 0.2547 |
| \( \partial S_4 \) | \( y = 0 \) | 0.5969 |
| \( \partial S_5 \) | \( x^2 + y^2 = 1 \) | 0.5969 |
Fig. 3: Probability density functions to generate data for (a) separable model, (b) nonseparable model.

Fig. 4: Boundaries \{\partial S_i\} considered in the separable model.

1) We first construct the indicator functions \(1_{S_i}\) and \(1_A\) on the corresponding graphs.

2) We then compute the empirical bandwidth \(\omega(1_{S_i})\) and \(\omega(1_A)\) in a manner that takes care of numerical error: we first obtain the eigenvectors of the corresponding \(L\), then set \(\omega(1_{S_i})\) and \(\omega(1_A)\) to be \(\nu\) for which energy contained in the Fourier coefficients corresponding to eigenvalues \(\lambda_j > \nu\) is at most 0.01%, i.e.,

\[
\omega(1_{S_i}) = \min \left\{\nu \mid \sum_{j : \lambda_j > \nu} (u_j^T 1_{S_i})^2 \leq 10^{-4}\right\} \quad (58)
\]

\[
\omega(1_A) = \min \left\{\nu \mid \sum_{j : \lambda_j > \nu} (u_j^T 1_A)^2 \leq 10^{-4}\right\}. \quad (59)
\]

The procedure above is repeated 100 times and the mean of \(\omega(1_{S_i})\) and \(\omega(1_A)\) are compared with \(\sup_{s \in \partial S_i} p(s)\) and \(\sup_{x \in \partial A} p(x)\) respectively. The result is plotted in Figure 5.

We observe that the empirical bandwidth is close to the theoretically predicted value and has a very low standard deviation. This supports our conjecture that stochastic convergence should hold for the bandwidth. To further justify this claim, we study the behavior of the standard deviation of \(\omega(1_{S_i})\).

Fig. 5: Convergence of empirical value of bandwidths \(\omega(1_{S_i})\) and \(\omega(1_A)\) for different boundaries \{\partial S_i\} and \(\partial A\) on corresponding graphs. Dark shaded regions denote standard deviation over 100 experiments. Red bars indicate theoretical values.

Fig. 6: Standard deviation of \(\omega(1_{S_i})\) and \(\omega(1_A)\) as a function of \(n\).

Fig. 7: Mean reconstruction errors averaged over 100 experiments for (a) \(1_{S_i}\), and (b) \(1_A\). Red-dashed lines indicate the theoretical label complexities of \(\partial S_3\) and \(\partial A\).
and $\omega(1,\epsilon)$ as a function of $n$ in Figure [4] where we observe a decreasing trend consistent with our result.

For our second experiment, we validate the bound on the label complexity of graph-based semi-supervised learning in Theorem [5] by reconstructing the indicator function corresponding to $\partial S_3$ and $\partial A$ from a fraction of labeled examples on the corresponding graphs. This is carried out as follows: For a given budget $B$, we find the set of points to label pivoted column-wise Gaussian elimination on the eigenvector matrix $\mathbf{U}$ of $\mathbf{L}$. This method ensures that the obtained labeled set guarantees perfect recovery for signals spanned by the first $B$ eigenvectors of $\mathbf{L}$ [12]. We then recover the indicator functions from these labeled sets by solving the least squares problem in [7] followed by thresholding. The mean reconstruction error is defined as

$$E_{\text{mean}} = \frac{\text{No. of mismatches on unlabeled set}}{\text{Size of unlabeled set}}. \tag{60}$$

We repeat the experiment 100 times by generating different graphs and plot the averaged $E_{\text{mean}}$ against the fraction of labeled examples. The result is illustrated in Figure [7]. We observe that the error goes to zero as the fraction of labeled points goes beyond the theoretically predicted label complexity as predicted. This reinforces the intuition that the bandwidth of class indicators is closely linked with the inherent geometry of the data.

VII. DISCUSSIONS AND FUTURE WORK

In this paper, we provided an interpretation of the graph sampling theoretic approach to semi-supervised learning. Our work analyzed the bandwidth of class indicator signals with respect to the Laplacian eigenvector basis and revealed its connection to the underlying geometry of the dataset. This connection is useful in justifying graph-based approaches for semi-supervised and unsupervised learning problems, and provides a geometrical interpretation of the smoothness assumptions imposed in the bandlimited reconstruction approach. Specifically, our results have shown that an estimate of the bandwidth of class indicators converges to the supremum of the probability density on the class boundaries for the separable model, and on the overlap regions for the nonseparable model. This quantifies the connection between the assumptions of smoothness (in terms of bandlimitedness) and low density separation, since boundaries passing through regions of low data density result in lower bandwidth of the class indicator signals. We numerically validated these results through various experiments.

There are several directions in which our results can be extended. In this paper we only considered Gaussian-weighted graphs, an immediate extension would be to consider arbitrary kernel functions for computing graph weights, or density dependent edge-connections such as $k$-nearest neighbors. Another possibility is to consider data defined on a subset of the $d$-dimensional Euclidean space.

Our analysis also sheds light on the label complexity of graph-based semi-supervised learning problems. We showed that perfect prediction from a few labeled examples using a graph-based bandlimited interpolation approach requires the same amount of labeling as one would need to completely encompass the boundary or region of ambiguity. This indicates that graph-based approaches achieve the theoretical label complexity as dictated by the underlying geometry of the problem. We believe that the main potential of graph-based methods will be apparent in situations where one can tolerate a certain amount of prediction error, in which case such approaches shall require fewer labeled data. We plan to investigate this as part of future work.

REFERENCES

[1] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. Semi-Supervised Learning (Adaptive Computation and Machine Learning). The MIT Press, 2006.

[2] Denny Zhou, Olivier Bousquet, Thomas N. Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In S. Thrun, L. K. Saul, and B. Schölkopf, editors, Advances in Neural Information Processing Systems 16, pages 321–328. MIT Press, 2004.

[3] Hariharan Narayanan, Mikhail Belkin, and Partha Niyogi. On the relation between low density separation, spectral clustering and graph cuts. In Advances in Neural Information Processing Systems (NIPS) 19, 2006.

[4] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In IN ICML, pages 912–919, 2003.

[5] Xueyuan Zhou and Mikhail Belkin. Semi-supervised learning by higher order regularization. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011, pages 892–901, 2011.

[6] Mikhail Belkin and Partha Niyogi. Using manifold structure for partially labeled classification. In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems 15, pages 953–960. MIT Press, 2003.

[7] Mikhail Belkin and Partha Niyogi. Semi-supervised learning on riemannian manifolds. Machine Learning, 56(1):209–239, 2004.

[8] I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Processing Magazine, 30(3):83–98, May 2013.

[9] A Anis, A Gadde, and A Ortega. Towards a sampling theorem for signals on arbitrary graphs. In Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, pages 3804–3808, May 2014.

[10] H. Shomorony and A.S. Avestimehr. Sampling large data on graphs. In Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on, pages 933–936, Dec 2014.

[11] S. Chen, R. Varma, A. Sandryhaila, and J. Kovac̆evi´c. Discrete signal processing on graphs: Sampling theory. Signal Processing, IEEE Transactions on, 2015.

[12] A. Anis, A. Gadde, and A. Ortega. Efficient sampling set selection for bandlimited graph signals using graph spectral proxies. IEEE Transactions on Signal Processing, 64(14):3775–3789, July 2016.

[13] Akshay Gadde, Aamir Anis, and Antonio Ortega. Active semi-supervised learning using sampling theory for graph signals. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14, pages 492–501, New York, NY, USA, 2014. ACM.

[14] Olivier Bousquet, Olivier Chapelle, and Matthias Hein. Measure based regularization. In Advances in Neural Information Processing Systems (NIPS) 16, MIT Press, 2004.

[15] Matthias Hein. Uniform Convergence of Adaptive Graph-Based Regularization, pages 50–64. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[16] Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for laplacian-based manifold methods. J. Comput. Syst. Sci., 74(8):1289–1308, 2008.

[17] Xueyuan Zhou, Mikhail Belkin, and Nathan Srebro. An iterated graph laplacian approach for ranking on manifolds. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011, pages 877–885, 2011.
2) We fill the locations one-by-one from left to right, using
1) All product terms have a form defined by the following
expansions of the individual product terms
summation of all possible
matrices in the product. Note that
Probability and Measure
Geometrical aspects of statistical learning theory
[20] S.K. Narang, A Gadde, and A Ortega. Signal processing techniques for
interpolation in graph structured data. In Acoustics, Speech and Signal Processing (ICASSP), 2013
IEEE International Conference on, April 2015.
2) We fill the locations one-by-one from left to right, using the
following set of rules. Let $K(x_a, x_b)$ be the term preceding
$K(x_i, x_i)$. Then
• If $K(x_a, x_b)$ is contributed by $D$, then $*=a$.
• If $K(x_a, x_b)$ is contributed by $W$, then $*=b$.
3) Further, let $K(x_a, x_{m+1})$ be the term preceding $1_S(x_s)$. Then
• If $K(x_a, x_{m+1})$ is contributed by $D$, then $*=a$.
• If $K(x_a, x_{m+1})$ is contributed by $W$, then $*=m+1$.

APPENDIX A

EXPANSION OF $\frac{1}{n} \sum_1^n L^m 1_S$ AND $\mathbb{E} \left\{ \frac{1}{n} \sum_1^n L^m 1_S \right\}$

To expand $\frac{1}{n} \sum_1^n L^m 1_S$ in terms of the elements $w_{ij}$ of $W$, we first write the expression for each product term. Since $L^m = \frac{1}{m!} (D - W)^m$, there are $2^m$ such terms. Let us first introduce the following notation: $[D, W]_m$ denotes a product term containing the matrices $D$ and $W$, such that there are $m$ matrices in the product. Note that $L^m$ is essentially a summation of all possible $[D, W]_m$ with appropriate signs.

Now, the explicit expression for $\frac{1}{n} \sum_1^n [D, W]_m 1_S$ can be obtained using the following procedure:

1) All product terms have a form defined by the following template:

$$\frac{1}{n} [D, W]_m 1_S = \sum_{i_1, \ldots, i_m+1} (1_S)_{i_1} w_{i_1 i_2} w_{i_2 i_3} \ldots w_{i_m i_{m+1}} (1_S)_{i_{m+1}},$$

where the locations with * need to be filled with appropriate indices depending on the product term. Note that each $w_{ij}$ is contributed by either a $D$ or $W$ depending on its location in the expression.

2) We fill the locations one-by-one from left to right, using the following set of rules. Let $w_{ab}$ be the term preceding $w_{ac}$, then
• If $w_{ab}$ is contributed by $D$, then $*=a$.
• If $w_{ab}$ is contributed by $W$, then $*=b$.

3) Let $w_{aim+1}$ denote the term preceding $(1_S)_{i_{m+1}}$. Then, we have the following rule:
• If $w_{aim+1}$ is contributed by $D$, then $*=a$.
• If $w_{aim+1}$ is contributed by $W$, then $*=m+1$.

The expansion of $\frac{1}{n} \sum_1^n L^m 1_S$ can be found by summing up the expansions of the individual product terms $\frac{1}{n} [D, W]_m 1_S$.

Recalling that $w_{ij} = K(X_i, X_j)$, we conclude

$$\frac{1}{n} \sum_1^n L^m 1_S = \frac{1}{n^{m+1}} \sum_{i_1, \ldots, i_m+1} g(X_{i_1}, X_{i_2}, \ldots, X_{i_m+1}).$$

The expression for $\mathbb{E} \left\{ \frac{1}{n} \sum_1^n [D, W]_m 1_S \right\}$ can be evaluated in a similar fashion, except that the summations are replaced by integrals. We first evaluate the expected value of individual product terms $\mathbb{E} \left\{ \frac{1}{n} [D, W]_m 1_S \right\}$ by the following rules:

1) The template for the expected value of any product term can be expressed through the following template:

$$\mathbb{E} \left\{ \frac{1}{n} \sum_1^n [D, W]_m 1_S \right\} = \int \cdots \int (1_S(x_1) K(x_1, x_2) K(x_2, x_3) \ldots K(x_s, x_{s+1}) 1_S(x_s) p(x_1) dx_1 \ldots p(x_{s+1}) dx_{s+1},$$

where the locations with * need to be filled with appropriate indices depending on the product term. Once again, each $K(x_s, x_i)$ is contributed by either a $D$ or $W$.

2) We fill the locations one-by-one from left to right, using the following set of rules. Let $K(x_a, x_b)$ be the term preceding $K(x_i, x_i)$. Then
• If $K(x_a, x_b)$ is contributed by $D$, then $*=a$.
• If $K(x_a, x_b)$ is contributed by $W$, then $*=b$.

3) Further, let $K(x_a, x_{m+1})$ be the term preceding $1_S(x_s)$. Then
• If $K(x_a, x_{m+1})$ is contributed by $D$, then $*=a$.
• If $K(x_a, x_{m+1})$ is contributed by $W$, then $*=m+1$.

APPENDIX B

PROOF OF LEMMA 4

The key ingredient required for evaluating the integrals in Lemma 4 involves selecting a radius $R (< \tau)$ as a function of $\sigma$ that satisfies the following properties as $\sigma \to 0$:

1) $R \to 0$, 
2) $R/\sigma \to \infty$, 
3) $R^2/\sigma \to 0$, 
4) $\epsilon_R/\sigma \to 0$, where $\epsilon_R := \int_{|z| > R} K_{\sigma^2}(0, z) dz$.

A particular choice of $R$ is given by $R = \sqrt{d \sigma^2 \log(1/\sigma^2)}$. Note that $R \to 0$ as $\sigma \to 0$. Further,

$$\frac{R}{\sigma} = \sqrt{d \log(1/\sigma^2)},$$

$$\frac{R^2}{\sigma^2} = d \sigma \log(1/\sigma^2).$$

Hence, $R/\sigma \to \infty$ and $R^2/\sigma \to 0$ as $\sigma \to 0$. Additionally, substituting the expression for $R$ in the tail bound for the norm of a $d$-dimensional Gaussian vector gives us:

$$\frac{\epsilon_R}{\sigma} = \frac{1}{\sigma} \int_{|z| > R} K_{\sigma^2}(0, z) dz \leq \frac{1}{\sigma} \left( \frac{\sigma^2 d}{R^2} \right)^{-d/2} e^{-d \frac{R^2}{2 \sigma^2} + \frac{d}{2}} \leq \frac{1}{\sigma} (\sigma^2 d \log(1/\sigma^2))^{d/2}.$$
We now consider the proof of equation (35), let
\[ I := \frac{1}{\sigma} \int_{S_2} \int_{\mathbb{R}^d} K_{\sigma^2}(x_1, x_2) p^n(x_1) p^d(x_2) \, dx_1 \, dx_2. \]  

(67)

Further, let \([S_1]_R\) indicate a tubular region of thickness \(R\) adjacent to the boundary \(\partial S\) in \(S_1\), i.e., the set of points in \(S_1\) at a distance \(\leq R\) from the boundary. Then, we have
\[
I = \frac{1}{\sigma} \int_{[S_1]_R} p^n(x_1) \left( \int_{S_2} K_{\sigma^2}(x_1, x_2) p^d(x_2) \, dx_2 \right) \, dx_1 \\
+ \frac{1}{\sigma} \int_{[S_1]_R} p^n(x_1) \left( \int_{\mathbb{R}^d} K_{\sigma^2}(x_1, x_2) p^d(x_2) \, dx_2 \right) \, dx_1. \]

(68)

\(E_I\) is the error associated with approximating \(I\) by \(I_1\) and exhibits the following behavior

\[ \lim_{\sigma \to 0} E_I = 0. \]

**Proof.** Note that
\[
E_I \leq \frac{1}{\sigma} \left( \max_{\|x\| > R} \right) \int_{[S_1]_R} p^n(x_1) \left( \int_{S_2} K_{\sigma^2}(x_1, x_2) \, dx_2 \right) \, dx_1 \\
\leq \frac{1}{\sigma} \left( \max_{\|x\| > R} \right) \int_{[S_1]_R} p^n(x_1) \left( \int_{\mathbb{R}^d} K_{\sigma^2}(0, z) \, dz \right) \, dx_1 \\
= \frac{\epsilon}{\sigma} \left( \max_{\|x\| > R} \right) \int_{[S_1]_R} p^n(x_1) \, dx_1 \\
\leq \frac{\epsilon}{\sigma} \left( \max_{\|x\| > R} \right). \]

(69)

Using \(\lim_{\sigma \to \infty} \epsilon_R / \sigma = 0\), we get the desired result. \(\blacksquare\)

In order to analyze \(I_1\), we need to define certain geometrical constructions (illustrated in Figure 8) as follows:

**Definition 1.** 1) For each \(x_1 \in [S_1]_R\), we define a transformation of coordinates as:
\[
x_1 = s_1 + r_1 n(s_1), \tag{70}
\]
where \(s_1\) is the foot of the perpendicular dropped from \(x_1\) onto \(\partial S\), \(r_1\) is the distance between \(s_1\) and \(x_1\), and \(n(s_1)\) is the surface normal at \(s_1\) (towards the direction of \(x_1\)). Since the minimum radius of curvature of \(\partial S\) is \(\tau\) and \(R < \tau\), this mapping is injective.

2) For each \(s_1 \in \partial S\), let \(H^+_s\) denote the halfspace created by the plane tangent on \(s_1\) and on the side of \(S_2\). Similarly, let \(H^+_s\) denote the halfspace on the side of \(S_1\) that is, \(H^+_s = \mathbb{R}^d \setminus H^+_s\).

3) Let \(W^+_s(x)\) denote an infinite slab of thickness \(x\) tangent to \(\partial S\) at \(s_1\) and towards the side of \(S_2\). Let \(W^-_s(y)\) denote a similar slab of thickness \(y\) on the side of \(S_1\).

4) Finally, for any \(x\), let \(B(x, R)\) denote the Euclidean ball of radius \(R\) centered at \(x\).

We now consider \(I_1\), the main idea here is to approximate the integral over \(S_2\) by an integral over the halfspace \(H^+_s\). Hence, we have:
\[
I_1 = \frac{1}{\sigma} \int_{[S_1]_R} p^n(x_1) \int_{H^+_s} K_{\sigma^2}(x_1, x_2) p^d(x_2) \, dx_2 \, dx_1 \\
+ \frac{1}{\sigma} \int_{[S_1]_R} p^n(x_1) \int_{\mathbb{R}^d - H^+_s} K_{\sigma^2}(x_1, x_2) p^d(x_2) \, dx_2 \, dx_1. \tag{71}
\]

where \(E_2\) is the error associated with the approximation. Therefore, we have
\[
I = I_2 + E_2 + E_1. \tag{72}
\]

We now show that as \(\sigma \to 0\), \(I_2 \to \frac{1}{\sqrt{2\pi}} \int_{\partial S} P^{\alpha + \beta}(s) \, ds\), and \(E_2 \to 0\).

**Lemma 10.** \(\lim_{\sigma \to 0} I_2 = \frac{1}{\sqrt{2\pi}} \int_{\partial S} P^{\alpha + \beta}(s) \, ds\).

**Proof.** Using the change of coordinates \(x_1 = s_1 + r_1 n(s_1)\), we have
\[
I_2 = \frac{1}{\sigma} \int_{\partial S} \int_0^R p^n(s_1 + r_1 n(s_1)) \left( \int_{H^+_s} K_{\sigma^2}(s_1 + r_1 n(s_1), x_2) p^d(x_2) \, dx_2 \right) \, ds_1 \, dr_1, \tag{73}
\]
where \(J(s_1, r_1)\) denotes the Jacobian of the transformation. Now, an arc \(PQ\) of length \(ds\) at a distance \(r_1\) away from \(\partial S\) gets mapped to an arc \(PQ'\) on \(\partial S\) whose length lies in the interval \([ds(1 - \frac{r_1}{\tau}), ds(1 + \frac{r_1}{\tau})]\). Therefore, for all points within \([S_1]_R\), we have
\[
(1 - \frac{r_1}{\tau})^{d-1} \leq |J(s_1, r_1)| \leq \left(1 + \frac{r_1}{\tau}\right)^{d-1}. \tag{74}
\]
Further, since \(p(x)\) is Lipschitz continuous with constant \(L_p\), \(p^n(x)\) is also Lipschitz continuous with constant \(L_{p,\alpha}\). There-
fore, for any \( x_1 \in [S_1]_R \), we have \( p^\alpha(x_1) = p^\beta(s_1) + L_{p,\alpha} R \).
This leads to the following simplification for \( I_2^2 \):

\[
I_2 = (1 + O(R^{d-1})) \int_{\partial S} p^\alpha(s_1)I_3(s_1)ds_1
+ O(R^d) \int_{\partial S} I_3(s_1)ds_1,
\]
where we defined

\[
I_3(s_1) := \frac{1}{\sigma} \int_0^R \int_0^{\infty} K_{\alpha^2}(s_1 + r, n(s_1), x_2)p^\beta(x_2)dx_2dr_1.
\]

Note that every \( x_2 \in H^+_1 \) can be written as \( s_2 + r_2 n(s_2) \), where \( n(s_2) = -n(s_1) \). Hence, we get

\[
I_3(s_1) = \int_{R^{d-1}} \left( 2\pi \sigma^2 \right)^{-\frac{d-1}{2}} e^{-\frac{(s_2 - x_2)^2}{2\sigma^2}} p^\beta(s_2 - r_2 n(s_1))ds_2
\times \frac{1}{\sigma} \int_0^R \int_0^{\infty} e^{-\frac{2s_1}{\sigma^2}} dr_1dr_2
\]

\[
= \left( \int_{R^{d-1}} \left( 2\pi \sigma^2 \right)^{-\frac{d-1}{2}} e^{-\frac{(s_2 - x_2)^2}{2\sigma^2}} p^\beta(s_2)ds_2 + O(R) \right)
\times \frac{1}{\sigma} \int_0^R \int_0^{\infty} e^{-\frac{2s_1}{\sigma^2}} dr_1dr_2
\]

\[
= \left( p^\beta(s_1) + O(\sigma^2) + O(R) \right)
\times \frac{1}{\sigma} \int_0^R \int_0^{\infty} e^{-\frac{2s_1}{\sigma^2}} dr_1dr_2,
\]
where we used Lipschitz continuity of \( p^\beta(x) \) in the second equality and applied Lemma 12 to arrive at the last step. Further, using the definition of the \( Q \)-function and integration by parts, we note that

\[
\frac{1}{\sigma} \int_0^R \int_0^{\infty} e^{-\frac{2s_1}{\sigma^2}} dr_1dr_2
= \int_0^{R/\sigma} \int_0^{\infty} e^{-\frac{2s_1}{\sigma^2}} dx dy
= \int_0^{R/\sigma} Q(y)dy
= \left[ \frac{R}{\sigma} Q\left( \frac{R}{\sigma} \right) + \frac{1}{\sqrt{2\pi}} \left( 1 - e^{-R^2/2\sigma^2} \right) \right].
\]

Therefore,

\[
I_3(s_1) = \left( p^\beta(s_1) + O(\sigma^2) + O(R) \right)
\times \left[ \frac{R}{\sigma} Q\left( \frac{R}{\sigma} \right) + \frac{1}{\sqrt{2\pi}} \left( 1 - e^{-R^2/2\sigma^2} \right) \right].
\]

Combining (75) and (78) and using the fact that \( R/\sigma \to \infty \) as \( \sigma \to 0 \) (from the definition of \( R \)), we get

\[
\lim_{\sigma \to 0} I_2 = \frac{1}{\sqrt{2\pi}} \int_{\partial S} p^\alpha + p^\beta(s)ds,
\]
which concludes the proof.

We now consider the error term \( E_2 \) and prove the following result:

**Lemma 11.** \( \lim_{\sigma \to 0} E_2 = 0 \).

**Proof.** Let us first rewrite \( E_2 \) as follows:

\[
E_2 = \frac{1}{\sigma} \int_{|S_1|_R} p^\alpha(x_1)I_4(x_1)dx_1,
\]
where we defined

\[
I_4(x_1) := \int_{S_2 - H^+_1} K_{\sigma^2}(x_1, x_2)p^\beta(x_2)dx_2.
\]

The key idea is to lower and upper bound \( I_4(x_1) \) for all \( x_1 \) using worst case scenarios and evaluate the limits of the bounds. Note that \( I_4(x_1) \) is largest in magnitude when \( S_1 \) or \( S_2 \) is a sphere of radius \( \tau \), as illustrated in Figures 9a and 9b. We now make certain geometrical observations. For any \( x_1 = s_1 + r_1 n(s_1) \in [S_1]_R \), we observe from Figure 9b that

\[
I_4(x_1) \leq \int_{W_s^-} K_{\sigma^2}(x_1, x_2)p^\beta(x_2)dx_2
\]

\[
+ \int_{B(x_1, R')} K_{\sigma^2}(x_1, x_2)p^\beta(x_2)dx_2
\]

\[
\leq \int_{W_s^-} K_{\sigma^2}(x_1, x_2)p^\beta(x_2)dx_2 + p^\beta_{\max} \epsilon_R.
\]

where \( R' = \frac{R^2}{2\tau} \). Similarly, from Figure 9a, we observe that

\[
I_4(x_1) \geq - \left[ \int_{W_s^+} K_{\sigma^2}(x_1, x_2)p^\beta(x_2)dx_2 \right]
\]

\[
+ \int_{B(x_1, R')} K_{\sigma^2}(x_1, x_2)p^\beta(x_2)dx_2
\]

\[
\geq - \left[ \int_{W_s^+} K_{\sigma^2}(x_1, x_2)p^\beta(x_2)dx_2 \right]
\]

Substituting these in (80) and using a simplification similar to that of \( I_2 \) in (75), we get

\[
E_2 \leq \left( 1 + O(R^{d-1}) \right) \int_{\partial S} p^\alpha(s_1)I_5^-(s_1)ds_1
\]

\[
+ O(R^d) \int_{\partial S} I_5^-(s_1)ds_1 + \frac{\epsilon_R}{\sigma} p^\alpha + p^\beta_{\max},
\]

\[
E_2 \geq - \left( 1 + O(R^{d-1}) \right) \int_{\partial S} p^\alpha(s_1)I_5^+(s_1)ds_1
\]

\[
- O(R^d) \int_{\partial S} I_5^+(s_1)ds_1 - \frac{\epsilon_R}{\sigma} p^\alpha + p^\beta_{\max},
\]

where we defined

\[
I_5^-(s_1) := \frac{1}{\sigma} \int_0^R \int_{W_{s_1}^-(R')} K_{\sigma^2}(s_1 + r_1 n(s_1), x_2)p^\beta(x_2)dx_2dr_1,
\]

\[
I_5^+(s_1) := \frac{1}{\sigma} \int_0^R \int_{W_{s_1}^+(R')} K_{\sigma^2}(s_1 + r_1 n(s_1), x_2)p^\beta(x_2)dx_2dr_1.
\]
Similar to the evaluation of $I^\alpha_3(s_1)$ in (77), we have
\[
I^\alpha_3(s_1) = (p^\beta(s_1) + O(\sigma^2) + O(R)) \times \frac{1}{\sigma} \int_0^R \int_0^{R'} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r_1-r_2)^2}{2\sigma^2}} \, dr_1 dr_2,
\]
and
\[
I^\beta_3(s_1) = (p^\alpha(s_1) + O(\sigma^2) + O(R)) \times \frac{1}{\sigma} \int_0^R \int_0^{R'} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r_1-r_2)^2}{2\sigma^2}} \, dr_1 dr_2.
\]
We now evaluate the two 1-D integrals as follows:
\[
\frac{1}{\sigma} \int_0^R \int_0^{R'} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r_1-r_2)^2}{2\sigma^2}} \, dr_1 dr_2 = 0\]
\[
\int_0^R \int_0^{R'} \frac{1}{\sqrt{2\pi\sigma^2}} \, dx dy = \int_0^R Q(y) - Q(y + R') \, dy
\]
\[
= \int_0^R Q(y) dy + \int_0^{R'} Q(y) dy - \int_0^{R+R'} Q(y) dy
\]
\[
= \frac{R}{\sigma} Q\left(\frac{R}{\sigma}\right) + \frac{R'}{\sigma} Q\left(\frac{R'}{\sigma}\right) - \frac{R + R'}{\sigma} Q\left(\frac{R + R'}{\sigma}\right)
\]
\[
= \frac{1}{\sigma} \left(1 - e^{-\frac{R^2}{2\sigma^2}} - e^{-\frac{R'^2}{2\sigma^2}} + e^{-\frac{(R+R')^2}{2\sigma^2}}\right).
\]
Noting that as $\sigma \to 0$, $R/\sigma \to \infty$ and $R'/\sigma \to 0$, we conclude that $\lim_{\sigma \to 0} E_2 = 0$.

The proof of (36) proceeds in a similar fashion by approximating the inner integral using hyperplanes. Specifically, similar to the proof of (35), we can show that the integral on the left hand side can be written as $I + E$, where
\[
I := \frac{1}{\sigma} \int_{[S_1]_{\alpha}} \int_{H_{\alpha}^{s_1}} [K_{\alpha\sigma}(x_1, x_2) p^\alpha(x_1) p^\beta(x_2)
\]
\[- K_{\beta\sigma}(x_1, x_2) p^\beta(x_1) p^\beta(x_2)] dx_1 dx_2,
\]
and $E$ is the residual associated with the approximation that can be shown to go to zero as $\sigma \to 0$ (we skip this proof since it is quite similar to the analysis for (35)). In order to evaluate $I$, we perform a change of coordinates $x_1 = s_1 + r_1 n(s_1)$ as before to obtain
\[
I = \frac{1}{\sigma} \int_{\partial S} \int_0^R p^\alpha(s_1 + r_1 n(s_1))
\]
\[- p^\beta(s_1 + r_1 n(s_1))
\]
\[\left(\int_{H_{\alpha}^{s_1}} K_{\alpha\sigma}(s_1 + r_1 n(s_1), x_2) p^\alpha(x_2) dx_2\right)
\]
\[- K_{\beta\sigma}(s_1 + r_1 n(s_1), x_2) p^\beta(x_2) dx_2\]
\[\left(\int_{H_{\beta}^{s_1}} K_{\beta\sigma}(s_1 + r_1 n(s_1), x_2) p^\beta(x_2) dx_2\right)
\]
\[\left|\det J(s_1, r_1)\right| ds_1 dr_1
\]
\[
= \int_{\partial S} p^\alpha(s_1) I_{\alpha}(s_1) ds_1 - \int_{\partial S} p^\beta(s_1) I_{\beta}(s_1) ds_1 + O\left(R^d\right),
\]
where we defined
\[
I_{\alpha}(s_1) := \frac{1}{\sigma} \int_0^R \int_{H_{\alpha}^{s_1}} K_{\alpha\sigma}(s_1 + r_1 n(s_1), x_2) p^\beta(x_2) dx_2 dr_1,
\]
\[
I_{\beta}(s_1) := \frac{1}{\sigma} \int_0^R \int_{H_{\beta}^{s_1}} K_{\beta\sigma}(s_1 + r_1 n(s_1), x_2) p^\beta(x_2) dx_2 dr_1.
\]
By using a change of coordinates for $x_2$ similar to the steps in (77), we obtain

\[
I_{\beta}(s_1) = \left( p^\beta(s_1) + O(\sigma^2) + O(R) \right) \times \\
\frac{1}{\sigma} \int_0^R \int_0^\infty \frac{1}{\sqrt{2\pi a\sigma^2}} e^{-\frac{(r_1-r_2)^2}{2a^2}} \, dr_1 \, dr_2, 
\]

(92)

\[
I_{\beta'}(s_1) = \left( p^{\beta'}(s_1) + O(\sigma^2) + O(R) \right) \times \\
\frac{1}{\sigma} \int_0^R \int_0^\infty \frac{1}{\sqrt{2\pi b\sigma^2}} e^{-\frac{(r_1-r_2)^2}{2b^2}} \, dr_1 \, dr_2. 
\]

(93)

The 1-D integrals can be evaluated as follows:

\[
\frac{1}{\sigma} \int_0^R \int_0^\infty \frac{1}{\sqrt{2\pi a\sigma^2}} e^{-\frac{(r_1-r_2)^2}{2a^2}} \, dr_1 \, dr_2 \\
= \sqrt{a} \int_0^R \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-y)^2}{2}} \, dx \, dy \\
= \sqrt{a} \int_0^R \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-y)^2}{2}} \, dx \, dy \\
= \sqrt{a} \int_0^R (1 - Q(y)) \, dy \\
= \frac{R}{\sigma} - \frac{R}{\sigma} \left( \frac{R}{\sqrt{a}\sigma} \right) - \sqrt{a} \sqrt{\frac{2\pi}{2\pi}} \left( 1 - e^{-R^2/2a^2} \right) ,
\]

\[
\frac{1}{\sigma} \int_0^R \int_0^\infty \frac{1}{\sqrt{2\pi b\sigma^2}} e^{-\frac{(r_1-r_2)^2}{2b^2}} \, dr_1 \, dr_2 \\
= \frac{R}{\sigma} - \frac{R}{\sigma} \left( \frac{R}{\sqrt{b}\sigma} \right) - \sqrt{b} \sqrt{\frac{2\pi}{2\pi}} \left( 1 - e^{-R^2/2b^2} \right) .
\]

Using the fact that $\alpha + \beta = \alpha' + \beta' = \gamma$, and taking the limit $\sigma \to 0$ after putting everything together, we conclude

\[
\lim_{\sigma \to 0} I = \frac{\sqrt{b} - \sqrt{a}}{\sqrt{2\pi}} \int_{\partial S} p^\gamma(s) \, ds. 
\]

(94)