Poisson reweighted Laplacian uncertainty sampling for graph-based active learning*

Kevin Miller† and Jeff Calder‡

Abstract. We show that uncertainty sampling is sufficient to achieve exploration versus exploitation in graph-based active learning, as long as the measure of uncertainty properly aligns with the underlying model and the model properly reflects uncertainty in unexplored regions. In particular, we use a recently developed algorithm, Poisson ReWeighted Laplace Learning (PWLL) for the classifier and we introduce an acquisition function designed to measure uncertainty in this graph-based classifier that identifies unexplored regions of the data. We introduce a diagonal perturbation in PWLL which produces exponential localization of solutions, and controls the *exploration* versus *exploitation* tradeoff in active learning. We use the well-posed continuum limit of PWLL to rigorously analyze our method, and present experimental results on a number of graph-based image classification problems.

Key words. active learning, uncertainty sampling, graph Laplacian, continuum limit, partial differential equations

MSC codes. 35J15, 35J20, 68T05, 35Q68

1. Introduction.

Supervised machine learning algorithms rely on the ability to acquire an abundance of labeled data, or data with known labels (i.e., classifications). While unlabeled data—data *without* known labels—is ubiquitous in most applications of interest, obtaining labels for such training data can be costly. Semi-supervised learning (SSL) methods leverage unlabeled data to achieve an accurate classification with significantly fewer training points. Simultaneously, the choice of training points can significantly affect classifier performance, especially due to the limited size of the training set of labeled data in the case of SSL. Active learning seeks to judiciously select a limited number of *query points* from the unlabeled data that will inform the machine learning task at hand. These points are then labeled by an expert, or human in the loop, with the aim of significantly improving the classifier performance.

While there are various paradigms for active learning [60], we focus on *pool-based* active learning wherein an unlabeled pool of data is available at each iteration of the active learning process from which query points may be selected. This paradigm is the natural fit for applying active learning in conjunction with semi-supervised learning since the unlabeled pool is also used by the underlying semi-supervised learner. These query points are selected by optimizing an *acquisition function* over the discrete set of points available in the unlabeled pool of data. That is, if $\mathcal{U} \subset \mathcal{X}$ is the set of currently unlabeled points in a pool of data inputs $\mathcal{X} \subset \mathbb{R}^d$, then the active learning process at each iteration selects the next query point $x^* \in \mathcal{U}$ to be

*Submitted to the editors DATE. **Source code:** https://github.com/millerk22/rwll_active_learning

Funding: JC was supported by NSF grant DMS:1944925, the Alfred P. Sloan foundation, and a McKnight Presidential Fellowship

†The Oden Institute of Computational Engineering and Sciences, University of Texas, Austin, TX (ksmiller@utexas.edu).

‡Department of Mathematics, University of Minnesota, Twin Cities, MN (jwcalder@umn.edu).

the minimizer of a real-valued acquisition function

$$x^* = \operatorname*{argmin}_{x \in \mathcal{U}} \ \mathcal{A}(x),$$

where $\mathcal{A}$ can depend on the current state of labeled information (i.e., the labeled data $\mathcal{L} = \mathcal{X} \backslash \mathcal{U}$ and corresponding labels for points in $\mathcal{L}$).

The above process (policy) for selecting query points is *sequential* as only a single unlabeled point is chosen to be labeled at each iteration, as opposed to the *batch* active learning paradigm. In batch active learning, a set of query points $\mathcal{Q} \subset \mathcal{U}$ is chosen at each iteration. While this is an important extension of the sequential paradigm and is an active area of current research in the literature [30, 51, 59, 67], we focus on the sequential case in this work.

Acquisition functions for active learning have been introduced for various machine learning models, especially support vector machines [2, 42, 66], deep neural networks [30, 47, 59, 62, 63], and graph-based classifiers [41, 50, 51, 55, 57, 79]. We focus on graph-based classifiers for our underlying semi-supervised learning model due to their ability to capture clustering structure in data and their superior performance in the *low-label rate regime*—wherein the labeled data constitutes a very small fraction of the total amount of data. Most active learning methods for deep learning assume a moderate to large amount of initially labeled data to start the active learning process. While there is exciting progress in improving the low-label rate performance of deep semi-supervised learning [58, 65, 74] and few-shot learning [37, 72], we restrict the focus of this paper to well-established graph-based paradigms for this setting.

An important aspect of the application of active learning in real-world datasets is the inherent tradeoff between using active learning queries to either explore the given dataset or exploit the current classifier's inferred decision boundaries. This tradeoff is reminiscent of the similarly named "exploration versus exploitation" tradeoff in reinforcement learning. In active learning, it is important to thoroughly explore the dataset in the early stages, and exploit the classifier's information in later stages. Algorithms that exploit too quickly can fail to properly explore the dataset, potentially missing important information, while algorithms that do not exploit the classifier in later stages can fail to efficiently refine classifier decision boundaries.

In this work, we provide a simple, yet effective, acquisition function for use in graph-based active learning in the low-label rate regime that provides a natural transition between exploration and exploitation summarized in a single hyperparameter. We demonstrate through both numerical experiments and theoretical results that this acquisition function explores prior to exploitation. We prove theoretical guarantees on our method by analyzing the continuum limit partial differential equation (PDE) that serves as a proxy for the discrete, graph-based operator. This is a novel approach to providing sampling guarantees in graph-based active learning. We also provide experiments on a toy problem that illustrates our theoretical results and the importance of the exploration versus exploitation hyperparameter in our method.

1.1. Previous work. The theoretical foundations in active learning have mainly focused on proving sample-efficiency results for linearly-separable datasets—frequently restricted to the unit sphere [1, 22, 34]—for low-complexity function classes using disagreement or margin-based acquisition functions [1, 2, 35, 36]. These provide convenient bounds on the number of active learning choices necessary for the associated classifier to achieve (near) perfect classification on these datasets with simple geometry. In contrast, much of the focus for theoretical work

in graph-based active learning leverage assumptions on the clustering structure of the data that is assumed to be captured in the graph structure [21, 55], which sometimes is assumed to be hierarchical [18, 23, 24]. A central priority in this line of inquiry establishes guarantees that, given assumptions about the clustering structure of the observed dataset $\mathcal{X}$, the active learning method in question will query points from *all* clusters (i.e., ensure exploration). The low-label rate regime of active learning—the focus of this current work—is the natural setting for establishing such theoretical guarantees.

The graph Laplacian has been widely used for semi-supervised learning over the past two decades, starting with the seminal work on Laplace learning (or label propagation, see [78]), and continued in a number of subsequent works [3, 5, 6, 13, 17, 33, 49, 69, 76]. Graph Laplacians are also used in spectral clustering [68] and spectral based embeddings [4, 20, 25]. Laplace learning is the underlying model for a number of graph-based active learning methods [41, 43, 50, 79]. However, relatively little work has been done to provide theoretical guarantees for exploration of clustering structure in these methods. These works instead focus on designing acquisition functions to (approximately) reduce the empirical risk [43] or variance [41] of a corresponding Gaussian random field on the discrete graph structure. Other important works in active learning have focused primarily on improving the performance of deep neural networks via active learning with either (1) moderate to large amounts of labeled data available to the classifier [30, 77] or (2) coreset methods that are agnostic to the observed labels of the labeled data seen throughout the active learning process [59, 67]. Our current work is focused on the *low-label rate regime*, which is an arguably more fitting regime for semi-supervised and active learning. Furthermore, in contrast to coreset methods, our acquisition function directly depends on the observed classes of the labeled data.

Graph neural networks (GNN) [70, 75] are an important area of graph-based methods for machine learning, and various methods for active learning have been proposed [9, 31, 40, 73]. GNNs consider network graphs whose connectivity is a priori determined via metadata relevant to the task (e.g., co-authorship in citation networks) and then use the node-level features to learn representations and transformations of features for the learning task. In contrast, we consider similarity graphs where the connectivity structure is determined only by the node-level features and directly learn a node function on this graph structure.

Continuum limit analysis of graph-based methods has been an active area of research for providing rigorous analysis of graph-based learning [10, 11, 13, 15, 17, 27, 32, 38, 39, 64]. In this analysis, a discrete graph is viewed as a random geometric graph that is sampled from a density $\rho : \mathbb{R}^d \to \mathbb{R}_+$ defined in a high-dimensional space (possibly constrained to a manifold $\mathcal{M} \subset \mathbb{R}^d$ therein). The graph Laplacian matrix can be analyzed via its continuum-limit counterpart, which is a second-order density weighted diffusion operator (or a weighted Laplace-Beltrami operator when the data is sampled from a manifold). An important development relevant to the current work is the Properly Weighted Graph Laplacian [17], which reweights the graph in the Laplace learning model of [78] to correct for the degenerate behavior of Laplace learning in the extremely low-label rate regime. This provides the setting for a well-defined, properly scaled graph-based semi-supervised learning model that we use in our current work to provide rigorous bounds on the acquisition function values to control the exploration versus exploitation tradeoff.

In order to apply active learning in practice, it is essential to design computationally

efficient acquisition functions. Much of the current literature has sought to design more sophisticated methods that often have higher computational complexity (e.g., requiring the full inversion of the graph Laplacian matrix). Uncertainty sampling [60] is an example of a computationally efficient acquisition function since it only requires the output of the classifier on the unlabeled data. However, uncertainty sampling methods will often mainly select query points that concentrate along decision boundaries while ignoring large regions of the dataset that are distant from any labeled points. Phrased in the terminology of the exploration versus exploitation tradeoff in reinforcement learning, uncertainty sampling is often overly "exploitative" and often achieves poor overall accuracy in empirical experiments [41].

In contrast, methods such as variance optimization (VOpt) [41], Σ-Opt [50], Coresets [59], LAND [55], and CAL [18] could be characterized as primarily "explorative" methods. Oftentimes, however, such explorative methods, or other methods that are designed to both explore and exploit [30, 44, 51, 79], are more expensive to compute than uncertainty sampling. For example, VOpt [41] and Σ-Opt [50] require the computation and storage of a dense $N \times N$ covariance matrix that must be updated after each active learning iteration. The work of [51] proposed a computationally efficient adaptation of these methods via a projection onto a subset of the graph Laplacian's eigenvectors. As a consequence of sometimes significantly poor performance from this spectral truncation method in our experiments, we provide a "full" computation of VOpt and Σ-Opt in certain experiments by restricting the computation to only a subset of unlabeled data which allows us to bypass the need to invert the graph Laplacian matrix (Section 3.4). This heuristic, however, is still very expensive to compute at each iteration making it not a viable option for moderate to large datasets in practice.

In this work, we show that uncertainty sampling, *when properly designed for the graph-based semi-supervised learning model* can both explore and exploit, and outperforms existing methods in terms of computational complexity, overall accuracy, and exploration rates.

1.2. Overview of paper. The rest of the paper continues as follows. We begin in Section 2 with a description of the Properly Weighted Laplace learning model from [17] that will be the underlying graph-based semi-supervised learning model for our proposed active learning method. We also introduce the minimum norm acquisition function in this section, along with other useful preliminaries for the rest of the paper. In Section 3, we begin with illustrative experiments in two-dimensions to illustrate the delicate balance between exploration and exploitation in graph-based active learning. Section 3.4 compares our proposed active learning method to other acquisition functions on larger, more "real-world" datasets that have been adapted to provide an experimental setup wherein exploration is essential for success in the active learning task. Thereafter, we present theoretical guarantees for the minimum norm acquisition function in the continuum limit setting in Section 4, along with an extended look at the theory in one dimension in Section 4.1.

1.3. Notation. Let $\|\cdot\|_2$ denote the standard Euclidean norm where the space is inferred from the input. We let $|\cdot|$ denote either the absolute value of a scalar in $\mathbb{R}$ or the cardinality of a set, where from context the intended usage should be clear. We denote the set of points $x \in \mathcal{X}$ with $x \notin \mathcal{U}$ as $\mathcal{X} \setminus \mathcal{U}$. We denote by $B_r(x) \subset \mathbb{R}^d$ the open ball of radius $r > 0$ centered at $x \in \mathbb{R}^d$ and write $B_r = B_r(0)$.

2. Model setup and acquisition function introduction. Let $\mathcal{X} = \{x_1, x_2, \dots, x_N\} \subset \mathbb{R}^d$ be a set of inputs for which we assume each $x \in \mathcal{X}$ belongs to one of C classes. Suppose that we have access to a subset $\mathcal{L} \subset \mathcal{X}$ of labeled inputs (*labeled data*) for which we have observed the ground-truth classification $y(x) \in \{1, \dots, C\}$ for each $x \in \mathcal{L}$. The rest of the inputs, $\mathcal{U} := \mathcal{X} \setminus \mathcal{L}$, are termed the *unlabeled data* as no explicit observation of the underlying classification have been seen for $x \in \mathcal{U}$. The semi-supervised learning task is to use both $\mathcal{L}$ and $\mathcal{U}$, with the associated labels $\{y(x)\}_{x\in\mathcal{L}}$, to infer the classification of the points in $\mathcal{U}$.

Sequential active learning extends semi-supervised learning by selecting a sequence of *query points* $x_1^*, x_2^*, \dots$ as part of an iterative process that alternates between (1) calculating the semi-supervised classifier given the current labeled information and (2) selecting and subsequently labeling an unlabeled query point $x_n^* \in \mathcal{U}_n$, where $\mathcal{U}_n = \mathcal{X} \setminus \mathcal{L}_n = \mathcal{X} \setminus (\mathcal{L} \cup \{x_1^*, x_2^*, \dots, x_{n-1}^*\})$. Labeling a query point x_i^* consists of obtaining the corresponding label $y(x_n^*)$ and then adding x_n^* to the set of labeled data, $\mathcal{L}_n = \mathcal{L}_{n-1} \cup \{x_n^*\}$. To avoid this cumbersome notation, however, we will drop the explicit dependence of $\mathcal{U}_n, \mathcal{L}_n$ on the iteration n and simply refer to the unlabeled and labeled data at the *current* iteration as respectively $\mathcal{U}$ and $\mathcal{L}$.

Returning to the underlying semi-supervised learning problem, graph Laplacians have often been used to propagate labeled information from $\mathcal{L}$ to $\mathcal{U}$ [7, 8, 10, 13, 17, 61, 70, 78]. From the set of feature vectors $\mathcal{X}$, consider a similarity graph $G(\mathcal{X}, W)$ with weight matrix $w_{ij} = \kappa(x_i, x_j)$ that captures the similarity between inputs x_i, x_j for each pair of points in $\mathcal{X}$. We use $\mathcal{X}$ to denote both the set of feature vectors as well as the node set for the graph G to avoid introducing more notation. Laplace learning [78] is an important graph-based semi-supervised learning model for both this current work and many previous graph-based active learning works, and solves the constrained problem of identifying a graph function $u : \mathcal{X} \to \mathbb{R}^C$ via the minimization of

$$\min_{u:\mathcal{X}\to\mathbb{R}^d} \sum_{x_i,x_j\in\mathcal{X}} w_{ij}\|u(x_i) - u(x_j)\|_2^2 \tag{2.1}$$

$$\text{subject to } u(x) = e_{y(x)} \text{ for } x \in \mathcal{L}.$$

The vector $e_{y(x)} \in \mathbb{R}^C$ is the standard Euclidean basis vector in $\mathbb{R}^C$ whose entries are all 0 except the entry corresponding to the label $y(x) \in \{1, \dots, C\}$. The learned function u that minimizes (2.1) constitutes a harmonic extension of the given labels in $\mathcal{L}$ to the unlabeled data. For the classification task, the inferred classification of $x \in \mathcal{U}$ is then obtained by thresholding on the learned function's output at x, $u(x) \in \mathbb{R}^C$. That is, the inferred classification $\hat{y}(x)$ for $x \in \mathcal{U}$ is given by

$$\operatorname*{argmax}_{c\in\{1,2,\dots,C\}} u_c(x),$$

where $u_c(x)$ denotes the c^{th} entry of $u(x)$.

Various previous works [13,16,17,28,29,56,61] have shown that when the amount of labeled information is small compared to the size of the graph (i.e., the *low-label rate regime*), the performance of minimizers of (2.1) degrades substantially. The solution u becomes roughly constant with sharp spikes near the labeled set, and the classification tends to predict the same label for most data points. Of particular interest to the current work is the Properly Weighted Laplace learning work in [17], wherein a weighting $\gamma : \mathcal{X} \to \mathbb{R}_+$ that scales like

dist$(x, \mathcal{L})^{-\alpha}$ for $\alpha > d-2$ is used to reweight the edges in the graph to correct the singular behavior of solutions to (2.1). We use an improvement to the Properly Weighted Laplacian that is called Poisson ReWeighted Laplace Learning (PWLL) and will be described in detail in another paper [14]. PWLL performs semi-supervised learning by solving the problem

$$\min_{u:\mathcal{X}\to\mathbb{R}^d} \sum_{x_i,x_j\in\mathcal{X}} \gamma(x_i)\gamma(x_j)w_{ij}\|u(x_i)-u(x_j)\|_2^2 \tag{2.2}$$

$$\text{subject to } u(x) = e_{y(x)} \text{ for } x \in \mathcal{L},$$

where the reweighting function γ is computed by solving the graph Poisson equation

$$\sum_{x_j\in\mathcal{X}} w_{ij}(\gamma(x_i)-\gamma(x_j)) = \sum_{x_k\in\mathcal{L}} \left(\delta_{ik} - \frac{1}{N}\right) \quad \text{for all } x_i \in \mathcal{X}. \tag{2.3}$$

In the previous work on the Properly Weighted Laplacian [17], the weight γ was explicitly chosen to satisfy $\gamma(x) \sim \text{dist}(x, \mathcal{L})^{-\alpha}$, while in the PWLL, γ is learned from the data, making the method more adaptive with fewer hyperparameters. The motivation for the Poisson equation (2.3) is that the continuum version of this equation is related to the fundamental solution of Laplace's equation, which produces the correct scaling in γ near the labeled set.

The reason for using PWLL is that minimizers of (2.2) have a well-defined continuum limit in the case when the amount of labeled data is fixed and the number of nodes $|\mathcal{X}| = N \to \infty$. This will allow us to analyze the behavior of our proposed minimum norm acquisition function applied to the PWLL model in the continuum limit setting in Section 4.

2.1. Solution decay parameter. We introduce an adaptation of (2.2) that increases the decay rate of the corresponding solutions away from labeled points. Controlling this decay will prove to be crucial for ensuring that query points selected via our minimum norm acquisition function (Section 2.2) will explore the extent of the dataset prior to exploiting current classifier decision boundaries. Given $\tau \geq 0$, we consider solutions to the following variational problem

$$\min_{u:\mathcal{X}\to\mathbb{R}^d} \sum_{x_i,x_j\in\mathcal{X}} \gamma(x_i)\gamma(x_j)w_{ij}\|u(x_i)-u(x_j)\|_2^2 \;+\; \tau \sum_{x_i\in\mathcal{U}} \|u(x_i)\|_2^2 \tag{2.4}$$

$$\text{subject to } u(x) = e_{y(x)} \text{ for } x \in \mathcal{L}.$$

It is straightforward to see that for $\tau > 0$ the additional term in (2.4) encourages the solution u to have *smaller* values away from the labeled data, where the values are fixed. When $\tau = 0$, we recover (2.2). We will refer to this graph-based semi-supervised learning model as Poisson ReWeighted Laplace Learning with τ-Regularization (PWLL-τ).

To illustrate the role of the decay parameter, let us consider a simple one dimensional version of this problem in the continuum of the form

$$\min_u \int_a^b u'(x)^2 + \tau u(x)^2 \, dx,$$

where $[a, b]$ is the domain and the minimization would be restricted by some boundary conditions on u (i.e., on the labeled set). Minimizers of this problem satisfy the ordinary differential

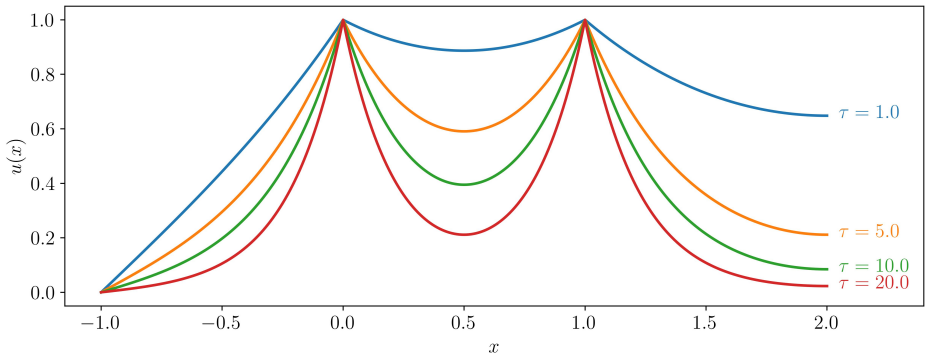


Figure 1. *Plots of solutions to $\tau u - u'' = 0$ for varying values of τ and with different boundary conditions. The intervals $(-1, 0), (0, 1), (1, 2)$ are the domains of three different solutions with boundary conditions $u(-1) = 0, u(0) = 1, u(1) = 1$, and $u'(2) = 0$. For increasing τ, the solutions decay more rapidly away from the points $x = -1, 0, 1$. This qualitative behavior is critical for demonstrating that our active learning acquisition function selects explorative query points in Section 4.*

equation (i.e., the Euler-Lagrange equation) $\tau u - u'' = 0$, which has two linearly independent solutions $e^{\pm\sqrt{\tau}x}$. Since the solution we are interested in is bounded, the exponentially growing one can be discarded, and we are left with exponential decay in the solutions with rate $\sqrt{\tau}$ away from the labeled set. In Figure 1, we plot a few example solutions for various values of τ and different boundary conditions to illustrate this exponential decay in one dimension. Thus, at least in this simple example, we can see how the introduction of the diagonal perturbation τ in PWLL leads to exponential decay of solutions, which is essential for the method to properly *explore* the dataset. We postpone developing this theory further until Section 4.

2.2. Minimum norm acquisition function. We now introduce the acquisition function that we propose to properly balance exploration and exploitation in graph-based active learning in the PWLL-τ model. We simply use the Euclidean norm of the output vector at each unlabeled point, $x \in \mathcal{U}$:

$$\mathcal{A}(x) = \|u(x)\|_2 = \sqrt{u_1^2(x) + u_2^2(x) + \ldots + u_C^2(x)}. \tag{2.5}$$

Due to the solution decay resulting from the τ-regularization term in (2.4), unlabeled points that are far from all labeled points will have small Euclidean norm (ℓ^2 norm) for their corresponding output vector. In the low-label rate regime, this property encourages query points selected by (2.5) to be spread out over the extent of the dataset, until a sufficient number of points have been labeled to "cover" the dataset. After this has been achieved in the active learning process, the learned functions for (2.4) will have smaller norms in regions between labeled points of differing classes due to the rapid decay in solutions near the transition between classes. This described behavior reflects the desired properties for balancing exploration prior to exploitation in active learning. Through both numerical experiments and theoretical results, we demonstrate this acquisition function's utility for this purpose.

The acquisition function (2.5) is a novel type of uncertainty sampling [60], wherein only the values of the learned function u at each active learning iteration are used to determine the

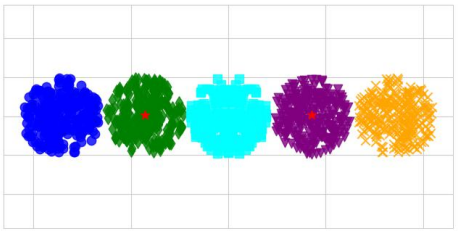
selection of query points. Note also that this acquisition function is *label adaptive* as opposed to *label agnostic*; that is, $\mathcal{A}(x)$ directly depends on the labelings of the currently labeled data, $\{y(x_j)\}_{x_j \in \mathcal{L}}$, since u does as well. Indeed, one may interpret the small Euclidean norm of the learned function at an unlabeled node, $\|u(x)\|_2$, to reflect uncertainty about the resulting inferred classification, $\hat{y}(x)$. Other uncertainty sampling methods, such as *smallest margin sampling* [60], also compute the uncertainty of the learned model at an unlabeled point via properties of the output vector $u(x) \in \mathbb{R}^C$. However, these criterion often either (1) only compare 2 entries of the vector to compute a measure of margin uncertainty or (2) normalize the output vector to lie on the simplex to be interpreted as class probabilities. In both cases, these measures of uncertainty in the classification of unlabeled points in unexplored regions of the dataset might not be as emphasized by the acquisition function compared to points that lie near the decision boundaries of the learned classifier. Our minimum norm acquisition function (2.5), however, is designed to prioritize the selection of query points in unexplored regions of the dataset which is properly reflected in the decay of the learned functions in the PWLL-τ model (2.4). In this sense, we are able to ensure exploration prior to exploitation in the active learning process using (2.5) in the PWLL-τ model.

Remark 2.1 (Choice of ℓ^2 norm). We briefly comment on the choice of ℓ^2 norm as the measure of uncertainty with the aid of an illustrative toy example. Consider the clustered dataset that is shown in Figure 2 (a), where we have distinguished the five different clusters with markers and colors. The assumed ground truth classification of the clusters is shown in panel (b), along with the initially labeled points plotted as red stars. Hence, with one initially labeled point in each class, we compute the PWLL-τ solution $u(x) = (u_1(x), u_2(x))^T \in \mathbb{R}^2$ for various values of $\tau \geq 0$ and plot the two components of $u_1(x), u_2(x)$ for each point in panels (c-e). The one-dimensional simplex is shown as a gray dotted line, and we see the effect of increasing $\tau > 0$ to "pull" points away from the simplex[1].

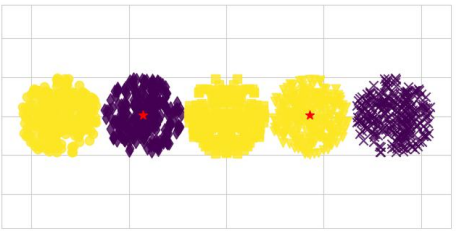
Since $u(x)$ values do not necessarily reside in the simplex for $\tau > 0$, we suggest that the ℓ^2 norm provides a useful measure of uncertainty that captures exploration. While some measures of uncertainty (e.g., entropy [60]) require the mapping of output values to the simplex, the ℓ^2 norm has no such requirement and consequently can differentiate between points lying between oppositely labeled points (e.g., cyan squares) and those residing in unexplored clusters (e.g., blue circles and orange x's). Further, once enough points have been labeled then values of $u(x)$ will lie relatively close to the simplex reflecting a transition from exploration to exploitation; the ℓ^2 norm values will then align with other traditional notions of uncertainty in active learning that are defined on the simplex ~~does~~. In contrast, a vector norm such as the ℓ^1 norm does not distinguish between points along the simplex and therefore would not lead to this natural transition from exploration to exploitation.

Remark 2.2 (Default to exploitation). In our PWLL-τ graph-based classifier, points whose outputs $u(x)$ lie near the center of simplex reside in regions of the domain between labeled points of differing labels (e.g., the cyan cluster of Figure 2). Thus, a consequence of using the

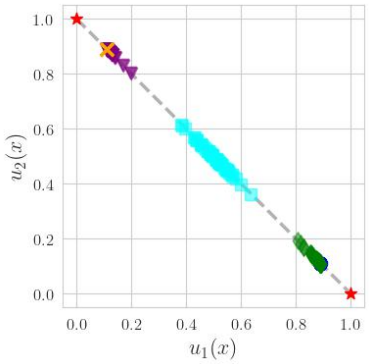
[1]Indeed when $\tau = 0$ the vector $u(x)$ is guaranteed to lie on the simplex due to the fact that the null space of the combinatorial (unnormalized) graph Laplacian matrix L for a connected graph is the span of constant vectors.



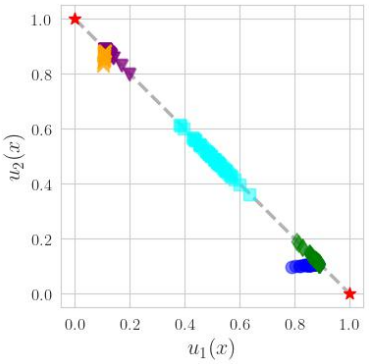
(a) Clusters and Init. Labeled



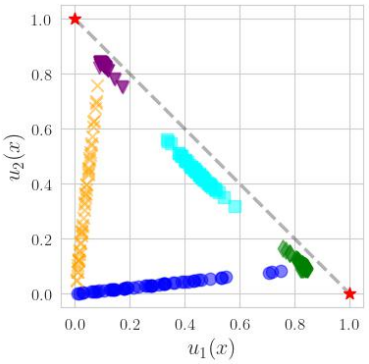
(b) Ground Truth Classification



(c) $\tau = 0$



(d) $\tau = 10^{-9}$



(e) $u(x) \in \mathbb{R}^2, \tau = 10^{-7}$

Figure 2. *Demonstration of utility of ℓ^2 norm for measure of uncertainty in PWLL-τ model. Panel (a) shows clusters identified by different colors and markers, with initially labeled points shown as red stars. Panel (b) shows the ground truth classification structure, and panels (c-e) show the output values of the PWLL-τ function $u(x) \in \mathbb{R}^2$ as they relate to the simplex (shown in gray dotted line). As $\tau > 0$ increases, the effect is that $u(x)$ in the outer clusters (blue circles and orange x's) is smaller and the ℓ^2 norm captures this effect.*

ℓ^2 norm to measure uncertainty reflects a "default to exploitation" since this favors selecting points whose output values lie closest to the center of the simplex. Our theoretical results in Section 4 focus accordingly on relating the value of $\tau > 0$ to the geometry of the dataset in order to guarantee cluster exploration when using this proposed acquisition function.

Remark 2.3 (Decay Schedule for τ). As we demonstrate through some toy experiments in Section 3.2, there is a benefit to decreasing the value of $\tau \geq 0$ as the active learning process progresses in order to more effectively transition from explorative to exploitative queries. While there are various ways to design this, we simply identify a constant $\mu \in (0, 1)$ so that the decreasing sequence of hyperparameter values $\tau_{n+1} = \mu\tau_n$ that satisfies $\tau_{2K} \leq \varepsilon$ with initial value $\tau_0 > 0$, where ε is chosen to be $\varepsilon = 10^{-9}$. For our experiments, we set K to be the number of clusters, which in the case of our tests is known a priori. In practice, this choice of K would be a user-defined choice to control the "aggressiveness" of the decay schedule of τ.

For $n \geq 2K$, we set $\tau_n = 0$. Thus, we calculate

$$\mu = \left(\frac{\varepsilon}{\tau_0}\right)^{\frac{1}{2K}} \in (0,1)$$

which ensures a decaying sequence of τ values as desired. We note that an interesting line of inquiry for future research would be to investigate a more rigorous understanding of how to adaptively select $\tau \geq 0$ during the active learning process. We leave this question for future research and simply use the proposed decay schedule above.

In Table 1, we introduce the abbreviations for and other useful information pertaining to the uncertainty sampling acquisition functions that we will consider in the current work—smallest margin, minimum norm, and minimum norm with τ-decay uncertainty sampling.

Full Name	**Abbreviation**	$\mathcal{A}(x)$	**Underlying Classifier**
Smallest Margin Unc. Sampling	Unc. (SM)	$u_{c_1^*}(x) - u_{c_2^*}(x)$	PWLL
Minimum Norm Unc. Sampling	Unc. (Norm)	$\|u(x)\|_2$	PWLL-τ, fixed $\tau > 0$
Minimum Norm Unc. Sampling with τ-decay	Unc. (Norm, $\tau \to 0$)	$\|u(x)\|_2$	PWLL-τ, decay $\tau \to 0$

Table 1

Description of uncertainty sampling acquisition functions that will be compared throughout the experiments in the following sections. Unc. (SM) considers the difference between the largest and second largest entries of the output vector $u(x)$, denoted by c_1^* and c_2^* respectively.

3. Results. In this section, we present numerical examples to demonstrate our claim that our proposed Unc. (Norm) and Unc. (Norm $\tau \to 0$) acquisition functions in the PWLL-τ model (2.4) are effective at both exploration and exploitation. We begin in Section 3.2 with a set of toy examples in 2-dimensions to facilitate visualizing the choices of query points during the active learning process and highlight the efficacy of implementing the τ-decay in Unc. (Norm, $\tau \to 0$) for balancing exploration and exploitation. In Section 3.3, we recreate an experiment from [41] on the Isolet dataset [26] to demonstrate that our proposed Unc. (Norm, $\tau \to 0$) essentially corrects previously observed negative behavior of uncertainty sampling.

In Section 3.4, we perform active learning experiments on larger, more "real-world" datasets. We use the **MNIST** [48], **FASHIONMNIST** [71], and **EMNIST** [19] datasets, and we interpret the original ground-truth classes (e.g. digits 0-9 in **MNIST**) as *clusters* on which we impose a different classification structure by grouping many clusters into a single class. This creates an experimental setting that necessitates exploration of initially unlabeled "clusters" in order to achieve high overall accuracy. We include similar experiments in Section **??** of the Supplemental Material to verify the performance of the proposed method in the presence of disparate class and cluster sizes.

While most previous work in the active learning literature (both graph-based and neural network classifiers) demonstrates acquisition function performance with only accuracy plots, we suggest another useful quantity for comparing performances. In the larger experiments of Sections 3.4 and **??**, we plot *the proportion of clusters that have been queried* as a function of

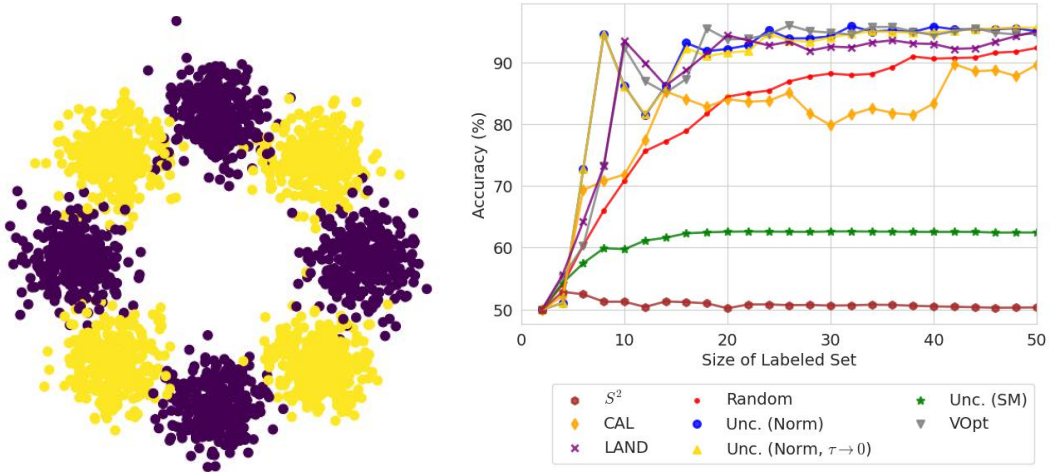
active learning iteration. These plots reflect how efficiently an acquisition function explores the clustering structure of the dataset, as captured by how quickly the proportionality curve increases toward 1.0. These cluster exploration plots are especially insightful for assessing performance in low label-rate active learning. An acquisition function that properly and consistently explores the clustering structure of the dataset will achieve an average cluster proportion of 1.0 faster than other acquisition functions and within a reasonable number of active learning queries.

3.1. Comparison to other methods. We comment here on a few notable methods in graph-based or geometry-inspired active learning that we include in some of our numerical comparisons: S^2 (Shortest-Shortest path) [21], LAND (Learning by Active Non-linear Diffusion) [55], and CAL (Cautious Active Learning) [18]. The S^2 algorithm by Dasarathy et al uses query points to bisect "shortest-shortest" paths in the graph between oppositely labeled points to recursively identify boundaries between clusters. While this method can efficiently sample query points along boundaries, S^2 essentially requires that initially labeled points belong to each of the respective clusters in the dataset. As such, it is admittedly at a disadvantage in a few of the experimental setups that we show herein. For example, the Isolet experiment 3.3 initially begins with only a single labeled point to test the explorative capabilities of the respective methods; in this experiment, we do not include a comparison to S^2 as it is not designed for such a setting.

In the LAND algorithm, Murphy and Maggioni use diffusion distances from a random walk interpretation of a similarity graph to select diverse sets of query points that are located in dense regions of the graph. Adjusting a model hyperparameter in the diffusion distances can reveal hierarchical clustering structure in the dataset which can encourage query points to be chosen at different resolution levels of the clustering structure. In a similar vein, the CAL algorithm by Cloninger and Mhaskar [18] uses hierarchical clustering structure to guide the query set selection process. By constructing a highly localized similarity kernel via Hermite polynomials, query points are selected at various resolution levels. Both the LAND and CAL algorithms have been shown to be effective at selecting query points in pixel classification for hyperspectral imagery applications. We, however, found that the current implementations of these algorithms were unable to scale to our larger experiments[2]. Hence, comparison to LAND and CAL are limited to the smaller experiments of Sections 3.2 and 3.3.

Furthermore, we suggest both the LAND and CAL methods may be more appropriately identified as "coreset" selection methods. Such methods leverage the geometry of the underlying dataset (e.g., the diffusion distances as captured by the similarity graph in LAND), but not the set of labels observed at labeled points during the active learning process. This is similar to other coreset methods that have been presented in both coreset and data summarization literature [54, 59, 67]. In contrast, our uncertainty-based criterion in this work combines both geometric information about the data as captured by the similarity graph structure and the observed labels at each labeled point via the output classification at each iteration. This makes our method more similar to the primary flavor of active learning methods.

[2]We adapted MATLAB implementations that were obtained from the respective authors for our experiments.



(a) Ground Truth (b) Accuracy Results

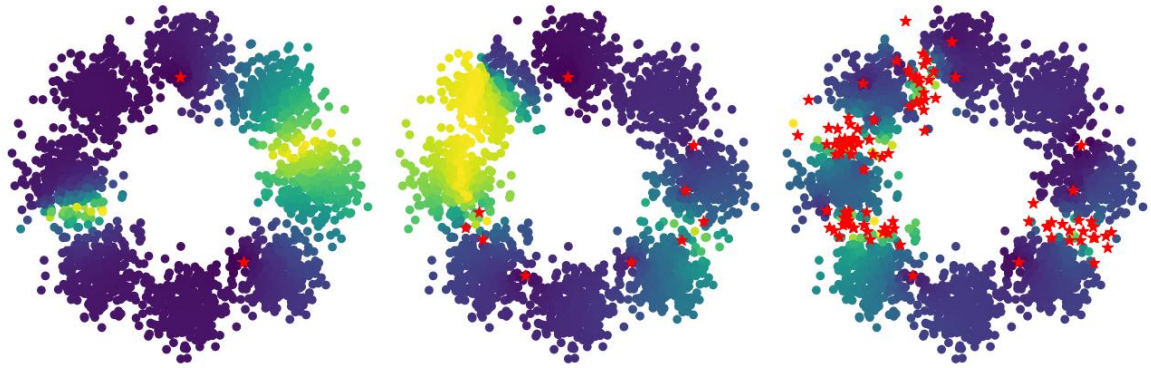
Figure 3. *Ground Truth (a) and Accuracy Results (b) for* ***Blobs*** *experiment. Notice that Unc. (SM) achieves very poor overall accuracy. We show in Figure 4 that this is due to premature exploitation.*

3.2. Toy examples. We first illustrate our claim regarding our minimum norm uncertainty sampling criterion for graph-based active learning with synthetic datasets that are directly visualizable (i.e., the data lies in only two dimensions). The first experiment—which we refer to as the **Blobs** experiment—illustrates how a non-zero value for τ in the initial phase of active learning is crucial for ensuring exploration of the dataset. The second experiment—which we refer to as the **Box** experiment—illustrates the need to decrease the value of τ to ensure the transition from exploration to exploitation. These experiments also allow us to directly observe the qualitative characteristics of the active learning query choices in uncertainty sampling.

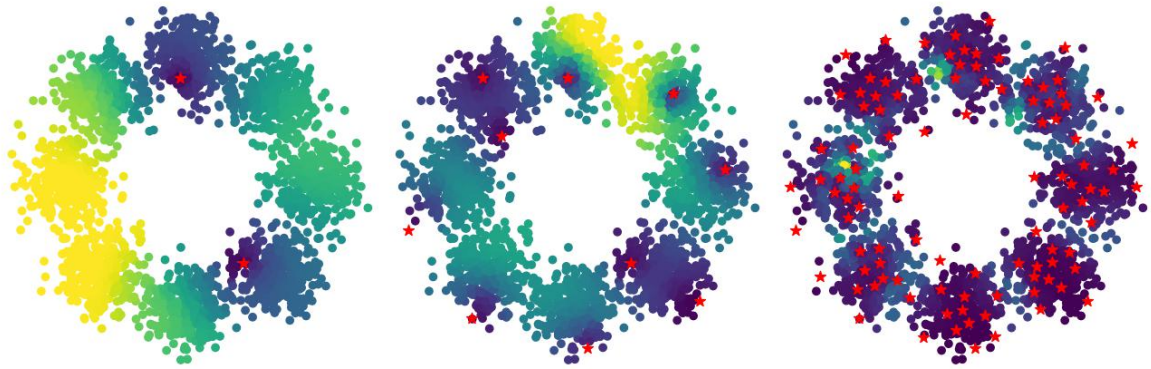
3.2.1. Blobs experiment. The **Blobs** dataset is comprised of eight Gaussian clusters, each of equivalent size (300) and variance ($\sigma^2 = 0.17^2$), whose centers (i.e., means) lie evenly spaced apart on the unit circle. That is, each cluster Ω_i is defined by randomly sampling 300 points from a Gaussian with mean $\mu_i = (\cos(\pi i/4), \sin(\pi i/4))^T \in \mathbb{R}^2$ and standard deviation $\sigma_i = \sigma = 0.17$. The classification structure of the clusters is then assigned in an alternating fashion, as shown in Figure 3(a). In each individual run of the experiment, one initially labeled point per *class* combine to be the starting labeled set, and then 100 active learning query points are selected sequentially via a specified acquisition function. Different acquisition functions then define different runs of the experiment.

For each acquisition function, we ran 10 experiments with different initially labeled points. The average accuracy at each iteration of an experiment is plotted in Figure 3(b). The main purpose of this experiment is to compare and contrast the characteristics of the query points selected by Unc. (SM), Unc. (Norm), and Unc. (Norm, $\tau \to 0$). For reference in these toy experiments, we include the results of using the VOpt [41] acquisition function as well as Random sampling (i.e., select $x_i^* \in \mathcal{U}$ with uniform probability over $\mathcal{U}$ at each iteration).

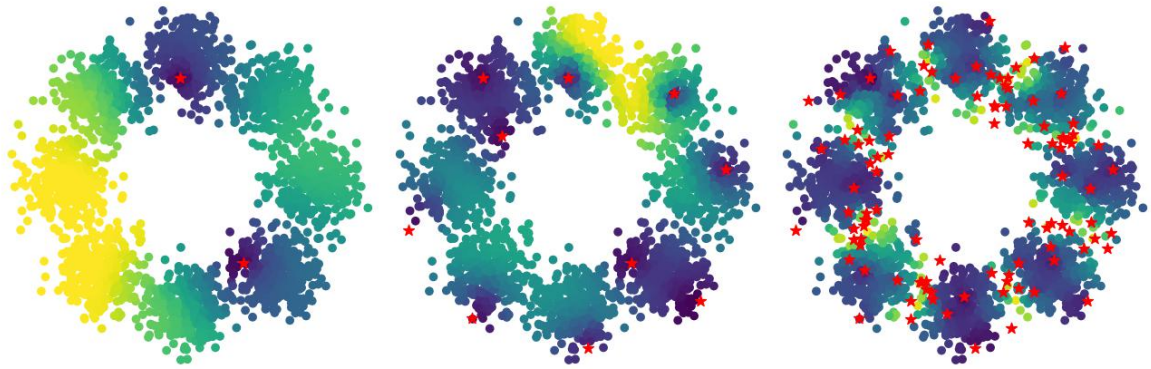
The main observation from this experiment is how poorly Unc. (SM) performs, as it only



(a) Unc. (SM), Initial (b) Unc. (SM), Iter 9 (c) Unc. (SM), Iter 100



(d) Unc. (Norm), Initial (e) Unc. (Norm), Iter 9 (f) Unc. (Norm), Iter 100



(g) Unc. (Norm, $\tau \to 0$), Initial (h) Unc. (Norm, $\tau \to 0$), Iter 9 (i) Unc. (Norm, $\tau \to 0$), Iter 100

Figure 4. *Acquisition Function Values for Unc. (SM), Unc. (Norm), and Unc. (Norm, $\tau \to 0$) at different stages of the* ***Blobs*** *experiment. Labeled points are marked as red stars and brighter regions of the heatmap indicate higher acquisition function values.*

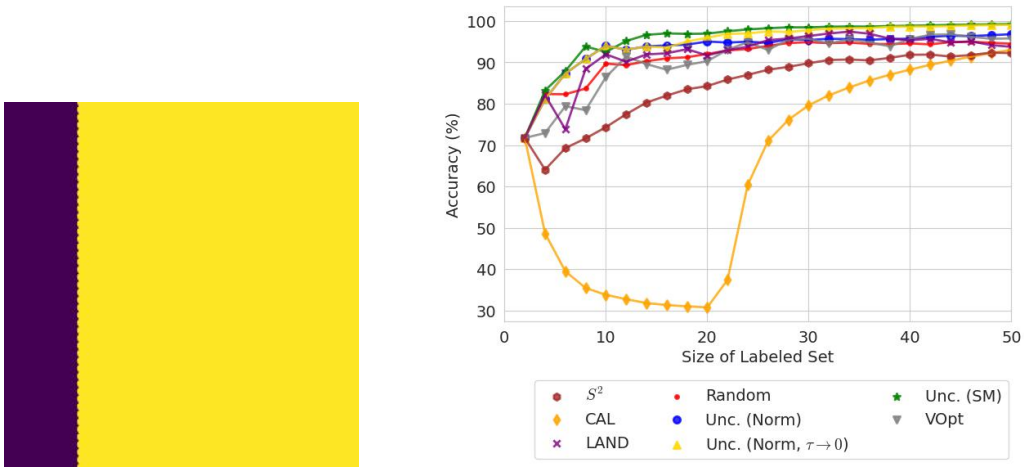
attains an overall accuracy of roughly 62% as the average over the trials. In Figure 4(a-c), we show one trial's acquisition function values heatmap at three different stages of the active learning process using Unc. (SM). We observe that the queries have primarily focused on the boundaries between a few clusters, while missing other clusters completely. At each iteration, the heatmap of acquisition function values has only focused on the current classifier's decision boundary which can lead to missing such clusters. In essence, we would qualify the behavior here as "premature exploitation", prior to proper exploration of the dataset.

In contrast, Figures 4 (d-i) demonstrate how the "minimum norm" uncertainty acquisition functions properly explore the extent of the geometric clustering structure. Both have sampled from every cluster in the ring. It is instructive to further see though that Unc. (Norm)—which employs a fixed value of $\tau > 0$ at every iteration—has not sampled more frequently *between* clusters by the end of the trial. We may characterize this behavior as not transitioning to proper exploitation of cluster boundaries. On the other hand, in Figure 4(i), we see that by using this minimum norm uncertainty sampling *with decaying values of* $\tau \to 0$ we more frequently sample at the proper cluster boundaries after having sampled from each cluster.

Remark 3.1. It is worth noting that, as an uncertainty sampling method that depends only on the current classifier's predictions, our Unc. (Norm) acquisition function does not take into account the influence that labeling a currently unlabeled point will have on the prediction of other points. This is in contrast to other more computationally intensive acquisition functions that explicitly model the "influence" that a currently unlabeled point has on the classifier's output predictions at other points (e.g., [41,43,44,50,51,79]). As such, our acquisition function may select query points that are not always ideal early on in the active learning process.

Consider, for example, panels (e) and (h) of Figure 4, wherein the selected query points by our acquisition functions lie in the outermost regions of the bottom clusters. These query points could constitute outliers and therefore not be the most influential on the classification of the other points in the corresponding clusters. While this behavior could severely hurt the performance of various classifiers, we note that the utilization of unlabeled data in graph-based learning through a similarity graph that explicitly models clustering structure helps to alleviate the potentially negative effects of such query points. As a result, the important behavior for our acquisition function is to first query points that belong to the different clusters (i.e., exploration) and then to query between oppositely labeled clusters (i.e., exploitation). Our empirical work suggests that the computational gains from using our inexpensive acquisition function are still meaningful despite the occasional selection of less influential query points.

3.2.2. Box experiment. The **Box** dataset is simply a 65×65 lattice of points on the unit square, with removing points that lie within a thin, vertical band centered at $x = 0.3$ which also defines the class boundary line (Figure 5). In contrast to the **Blobs** experiment, the **Box** experiment illustrates the need to transition from exploration to exploitation, and how this is accomplished by decreasing $\tau \to 0$. In the accuracy plot (Figure 5(b)), notice how the accuracy achieved by Unc. (Norm) levels off at a *lower* overall accuracy than both Unc. (SM) and Unc. (Norm $\tau \to 0$). Figure 6 demonstrates that this is due to "over exploration" of the dataset instead of transitioning to refining the decision boundary between classes. Active learning seeks to balance exploration versus exploitation while still being sample efficient,



(a) Ground Truth (b) Accuracy Results

Figure 5. *Ground Truth (a) and Accuracy Results (b) for* ***Box*** *experiment. Notice that Unc. (Norm) achieves suboptimal overall accuracy. We show in Figure 6(f) that the distribution of query points later in the active learning process reflect a lack of transition to exploitation.*

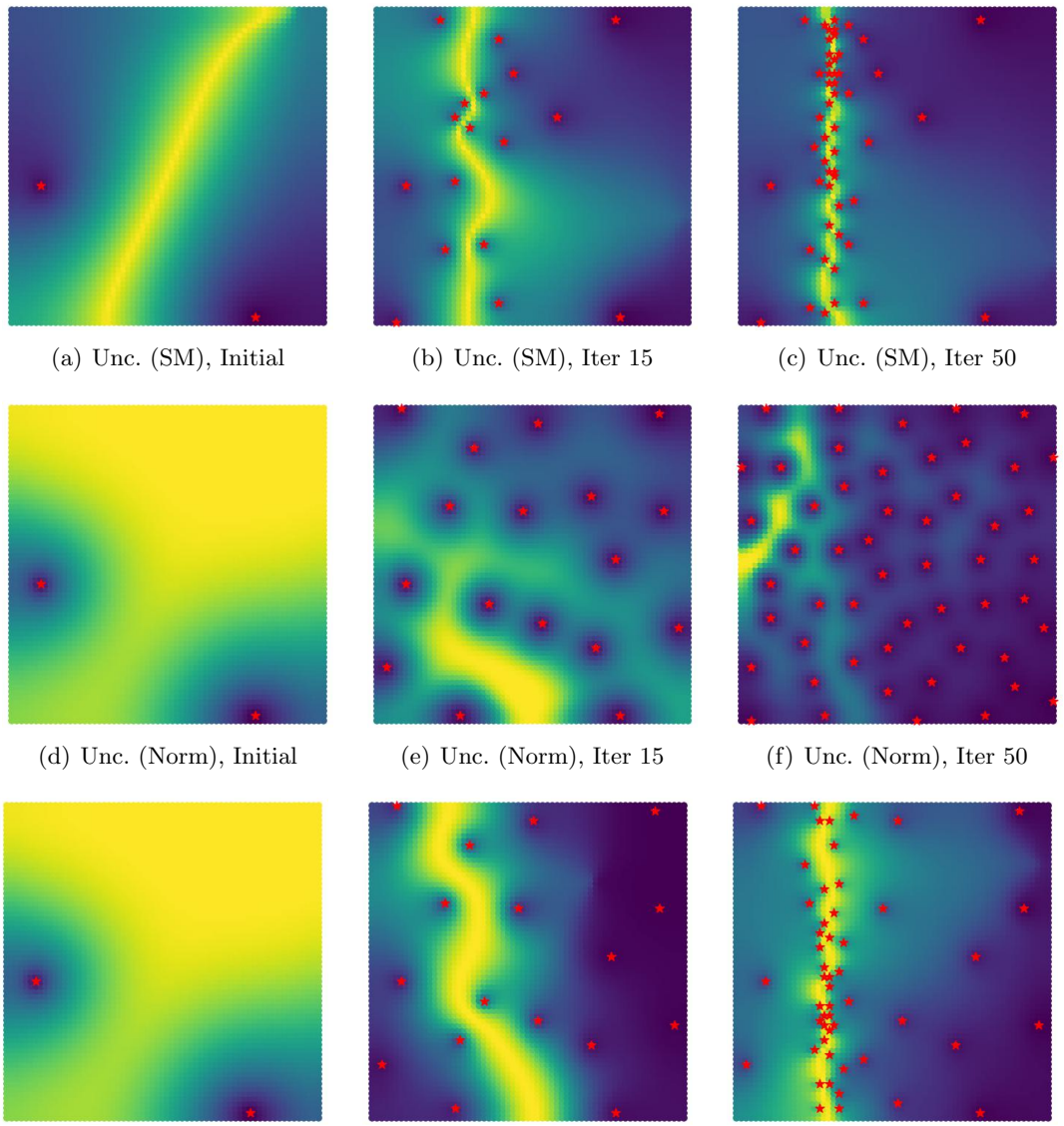
making as few active learning queries as possible.

As shown in Figures 6 (a-f), both Unc. (SM) and Unc. (Norm, $\tau \to 0$) more efficiently sample the decision boundary between the two classes in this **Box** dataset. Due to the very simple structure of the dataset, purely exploiting decision boundary information—as done by Unc. (SM)—is optimal. In contrast, Unc. (Norm, $\tau \to 0$) ensures to sparsely explore the extent of the right side of the box *prior to* exploiting the decision boundary. This is due to the decreasing value of τ over the iterations, and allows for a straightforward transition between exploration and exploitation. We set the value of $K = 8$ for the τ-decay schedule so that by 8 active learning queries we have transitioned to exploitation.

3.2.3. Overall observations. From the toy experiments presented in Sections 3.2.1 and 3.2.2, we see that the minimum norm uncertainty sampling *with decaying values of* τ has the desired behavior for a sample-efficient criterion that both explores and exploits during the active learning process. Ensuring this behavior in uncertainty sampling is also desirable because of the relatively light computational complexity that uncertainty sampling incurs. We now demonstrate on more complicated, "real-world" datasets the effectiveness of minimum norm uncertainty sampling in graph-based active learning.

3.3. Isolet case study. Our minimum norm uncertainty sampling in the PWLL-τ model can overcome previously negative results that have characterized uncertainty sampling. In [41], the authors introduced the Variance Optimization (i.e., VOpt) acquisition function and showcased this acquisition function on the Isolet spoken letter dataset[3] from the UCI repository [26], which contains 26 different classes. They compared against smallest margin uncertainty sampling (Unc. (SM)) among other acquisition functions. Of particular interest to us is how

[3]Accessed via https://archive.ics.uci.edu/ml/datasets/isolet.



(a) Unc. (SM), Initial (b) Unc. (SM), Iter 15 (c) Unc. (SM), Iter 50

(d) Unc. (Norm), Initial (e) Unc. (Norm), Iter 15 (f) Unc. (Norm), Iter 50

(g) Unc. (Norm, $\tau \to 0$), Initial (h) Unc. (Norm, $\tau \to 0$), Iter 15 (i) Unc. (Norm, $\tau \to 0$), Iter 50

Figure 6. *Acquisition Function Values for Unc. (SM), Unc. (Norm), and Unc. (Norm, $\tau \to 0$) at different stages of the* ***Box*** *experiment. Labeled points are marked as red stars and brighter regions of the heatmap indicate higher acquisition function values.*

poorly Unc. (SM) performed on this task, resulting in significantly worse accuracies than even random sampling.[4] In Supplemental Material Section **??**, we demonstrate that similar—even superior—performance is attained on this task by simply using our minimum norm uncertainty sampling (Unc. (Norm)). This highlights that our proposed uncertainty sampling method is

[4]We refer the reader to original paper [41] for more details.

more appropriate for low-label rate active learning than previous uncertainty sampling methods which have been characterized as overly-exploitative in the low-label rate regime. See Section **??** in the Supplemental Material for further details.

3.4. Larger datasets. In this section, we present the results of active learning experiments for multiclass classification problems derived from the **MNIST** [48], **FASHIONMNIST** [71], and **EMNIST** datasets [19]. We construct similarity graphs for each of these datasets by first embedding the points via the use of variational autoencoders (VAE) [45, 46] that were previously trained[5] in an unsupervised fashion, similar to [13].

Since a main crux of the present work is to ensure *both* exploration of clusters in a dataset and exploitation of cluster boundaries, we adapt the classification structure of the above datasets to require both. That is, we take the "true" class labelings $y_i \in \{0, 1, \dots, C\}$ (e.g. digits 0-9 for **MNIST**) and reassign them to one of $K < C$ classes by taking $y_i^{new} \equiv y_i \bmod K$; see Table 2 below.

Resulting Mod Class	0	1	2	3	4
MNIST	0,3,6,9	1,4,7	2,5,8	-	-
FASHIONMNIST	0,3,6,9	1,4,7	2,5,8	-	-
EMNIST	0,5,...,45	1,6,...,46	2,7,...,42	3,8,...,43	4,9,...,44

Table 2

Mapping of ground truth class label to $\bmod K$ *labeling for experiments of Section 3.4. Each ground truth class is interpreted as a different "cluster" and the resulting class structure for the experiments have multiple clusters per class. For* ***MNIST*** *and* ***FASHIONMNIST****, there are 10 ground truth classes and we take labels modulo* $K = 3$*. For* ***EMNIST****, there are 47 total ground truth classes and we take labels modulo* $K = 5$*.*

For each trial of an acquisition function, we select one initially labeled point per *"modulo"* class; therefore, only a subset of "clusters" (i.e., the original true classes) has an initially labeled point. In order to perform active learning successfully in these experiments, query points chosen by the acquisition function must sample from each cluster. In this way, we have created an experimental setup with commonly used machine learning datasets with potentially more complicated clustering structures wherein we test and compare the following acquisition functions: Uncertainty Sampling (SM), Unc. (Norm), Unc. (Norm, $\tau \to 0$), Random, VOpt [41] (see Remark 3.3), Σ-Opt [50] (also see Remark 3.3), and MCVOpt [52]. We perform 10 trials for each acquisition function, where each trial begins with a different initially labeled subset. To clarify, trials begin with only 3 labeled points in the **MNIST** and **FASHIONMNIST** experiments and with only 5 labeled points in the **EMNIST** experiments.

In the left panel of Figures 7-9, we show the accuracy performance of each acquisition function averaged over the 10 trials. The right panels of each of these figures display the average proportion of clusters that have been sampled by the acquisition functions at each

[5]The representations for **MNIST** and **FASHIONMNIST** are available in the GraphLearning package [12], while the code used to train the VAE for **EMNIST** is available in our Github repo https://github.com/millerk22/rwll_active_learning.

iteration of the active learning process. We refer to these plots as "Cluster Exploration" plots since they directly assess the explorative capabilities of the acquisition functions in question.

We observe that across these experiments, both Unc. (Norm) and Unc. (Norm, $\tau \to 0$) consistently achieve the best accuracy and cluster exploration results. It is somewhat surprising that without decaying τ, the Unc. (Norm) acquisition function seems to perform the best even after each cluster has been explored. The experiments in Section 3.2 suggest that the optimal performance in the exploitation phase of active learning would require taking $\tau \to 0$. We hypothesize that the clustering structure of relatively high-dimensional data—like these datasets—is much more complicated than our intuition would suggest from analyzing toy and other visualizable (i.e., 1D, 2D, or 3D) datasets. Regardless, we see that the minimum norm uncertainty acquisition function consistently outperforms other acquisition functions in these low-label rate active learning experiments.

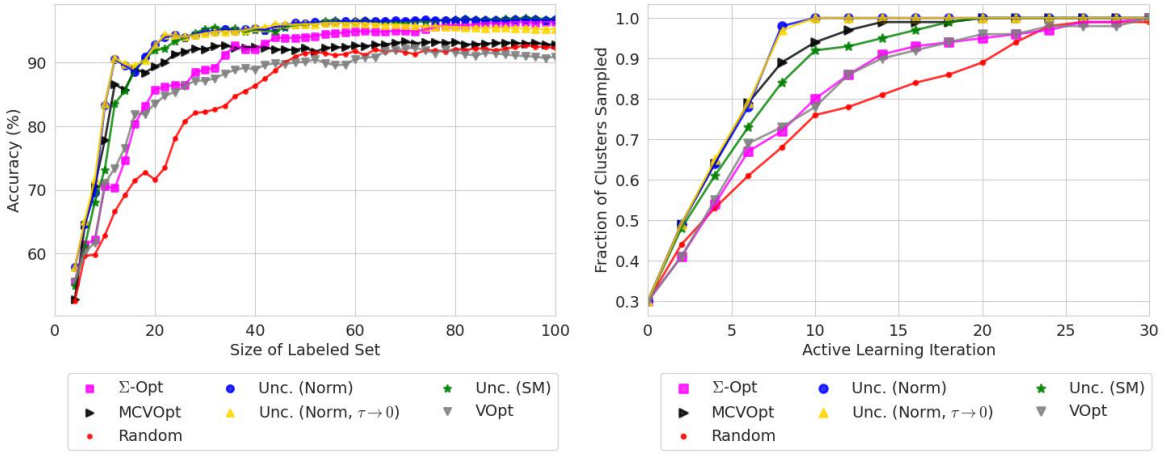
Remark 3.2 (Computational cost of Unc. (Norm)). While some acquisition functions such as VOpt, Σ-Opt, and MCVOpt require the computation, storage, and update of large auxiliary variables (e.g., inverse of graph Laplacian matrix), our proposed Unc. (Norm) acquisition function only requires the PWLL-τ solution with the currently labeled data in $\mathcal{L}$, which we compute with the preconditioned conjugate gradient method. Indeed, one of the reasons that we refer to our acquisition function as an uncertainty sampling criterion is that like previous uncertainty sampling methods [60], our acquisition function is simply a function of the current classifier. In this sense, Unc. (Norm) is "as cheap" as one could hope for in an acquisition function that depends on the labels of currently labeled data through the outputs of the underlying classifier. We have included Table 3 to compare computational costs among comparable methods; the S^2 [21], LAND [55], and CAL [18] are not included since they do not follow the same common framework of selecting $x^* = \operatorname{argmin}_{x \in \mathcal{U}} \mathcal{A}(x)$ at each iteration. The computational cost of solving the graph Laplace equation for training the model is omitted since it is shared by all algorithms (and is the only substantial cost with Unc. (Norm)).

Abbr. Name	Aux. Overhead	Cost Per Unlabeled	Aux. Update Cost
Unc. (Norm)	-	$\mathcal{O}(C)$	-
VOpt	$\mathcal{O}(N^3)$	$\mathcal{O}(N)$	$\mathcal{O}(N^2)$
Σ-Opt	$\mathcal{O}(N^3)$	$\mathcal{O}(N)$	$\mathcal{O}(N^2)$
MCVOpt	$\mathcal{O}(N^2 r)$	$\mathcal{O}(r + C)$	$\mathcal{O}(Nr)$

Table 3

Computational comparison between acquisition functions, where $N = |\mathcal{X}|$, C is the number of classes, and $r \ll N$ is the number of eigenvalues computed in the auxiliary matrix used in MCVOpt [53].

Remark 3.3. Due to the large nature of these datasets, computing the original VOpt and Σ-Opt criteria are inefficient (and often intractable) since this requires computing the inverse of a perturbed graph Laplacian matrix; this inverse is dense and burdensome to store in memory. We initially used an approximation that utilizes a subset of eigenvalues and eigenvectors of the graph Laplacian, similar to what was done in [51]. While this performed relatively well on the **EMNIST** experiment, we noticed significantly poor results on the **MNIST** and



(a) Accuracy (b) Cluster Proportion

Figure 7. *Accuracy Results (a) and Cluster Proportion (b) plots for **MNIST** dataset.*

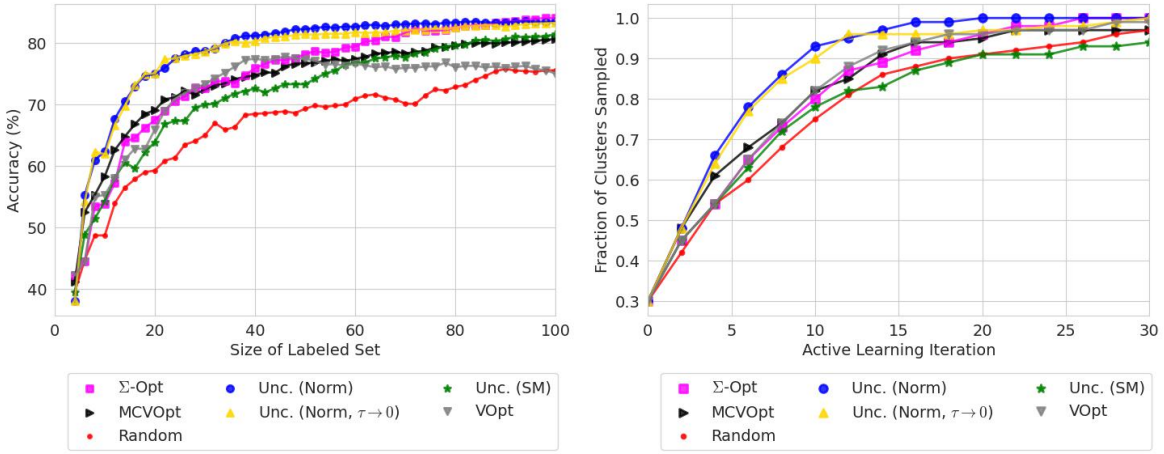
FASHIONMNIST experiments seemingly due to the spectral truncation with a resulting oversampling of a single cluster during the active learning process.

As an alternative to the spectral truncation, we performed a "full" calculation of these acquisition functions on a small, random subset of 500 unlabeled points at each active learning iteration. This performed significantly better than the spectral truncation in the **MNIST** and **FASHIONMNIST** experiments, and so we report these spectral truncation results in this section. In Figures 7 and 8 we refer to this simply by the original names, VOpt and Σ-Opt, respectively. The small choice of unlabeled points on which to evaluate the acquisition function is due to the burdensome computation needed at each step that scales with the size of this subset; at this reported choice of 500 points each active learning iteration already takes roughly 6 minutes to complete for the **MNIST** dataset. Due to its even greater size, we do not compute this "full" calculation on the random subset for the **EMNIST** dataset, but remark that the performance of the approximate (spectral truncation) VOpt and Σ-Opt already achieve comparable accuracy to the other reported methods in this dataset. We denote the spectral truncation with the suffix "(ST)" in Figure 9.

4. Continuum analysis of active learning. We now study our active learning approach rigorously through its continuum limit on an open, bounded set $\Omega \subset \mathbb{R}^d$ on which our data-points are sampled from. As was shown in [17], the continuum limit of (2.2) is the family of singularly weighted elliptic equations

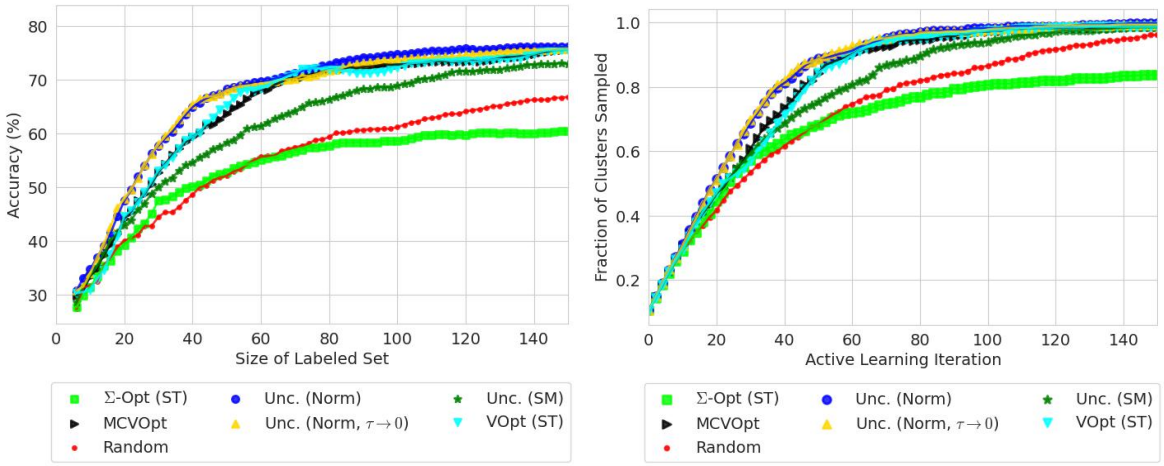
$$\begin{cases} \tau u_i - \rho^{-1} \operatorname{div}\left(\gamma \rho^2 \nabla u_i\right) = 0, & \text{in } \Omega \setminus \mathcal{L} \\ u_i = 1, & \text{on } \mathcal{L}_i \\ u_i = 0, & \text{on } \mathcal{L} \setminus \mathcal{L}_i, \end{cases} \tag{4.1}$$

where $\rho(x) \geq \rho_{min} > 0$ is the density of the data points, γ is the singular reweighting, described in more detail below, $\mathcal{L}_i \subset \Omega$ are the labeled points in the i^{th} class, and $\mathcal{L} = \cup_{i=1}^{C} \mathcal{L}_i$



(a) Accuracy (b) Cluster Proportion

Figure 8. *Accuracy Results (a) and Cluster Proportion (b) plots for* ***FASHIONMNIST*** *dataset.*



(a) Accuracy (b) Cluster Proportion

Figure 9. *Accuracy Results (a) and Cluster Proportion (b) plots for* ***EMNIST*** *dataset.*

the locations of all labeled points. The notation ∇ refers to the gradient vector and div is the divergence. We assume that the density $\rho(x) : \Omega \to (0, \infty)$ is Lipschitz continuous. The solutions u_i also satisfy the homogeneous Neumann boundary condition $\nabla u \cdot \nu = 0$ on $\partial\Omega$, where ν is the outward unit normal vector to Ω, but we omit writing this as it is not directly used in any of our arguments. We assume the sets $\mathcal{L}_i$ are all finite collections of points. The classification decision for any point $x \notin \mathcal{L}$ is given by

$$\ell(x) = \operatorname*{argmax}_{1 \le i \le C} u_i(x).$$

The continuum version of the uncertainty sampling acquisition function is then given by

$$\mathcal{A}(x) = \sqrt{u_1(x)^2 + u_2(x)^2 + \cdots + u_C(x)^2}. \tag{4.2}$$

As alluded to in Remark 2.2, the aim of this section is to use continuum PDE analysis to rigorously establish exploration guarantees in light of the exploitation default of uncertainty norm sampling (4.2), and illustrate how it depends on the choice of the decay parameter τ.

4.1. Illustrative 1D continuum analysis. We proceed first with an analysis of the continuum equations (4.1) in the one-dimensional setting, where the equations are ordinary differential equations (ODEs). The conclusions are insightful for the subsequent generalization to higher dimensions in Section 4.

Consider an interval $\Omega = (x_{min}, x_{max}) \subset \mathbb{R}$ with density $0 < \rho_{min} \leq \rho(x) \leq \rho_{max} < +\infty$. Assume a binary classification structure on this dataset, and further assume we have been given at least one labeled point per class. Let the pairs $\{(x_i, y_i)\}_{i=1}^{\ell} \subset \Omega \times \{1, 2\}$ be the input-class values for the currently labeled points ordered such that $x_i < x_{i+1}$. For ease in our discussion, we also assume that $x_1 = x_{min}$ and $x_\ell = x_{max}$ (Figure 10).

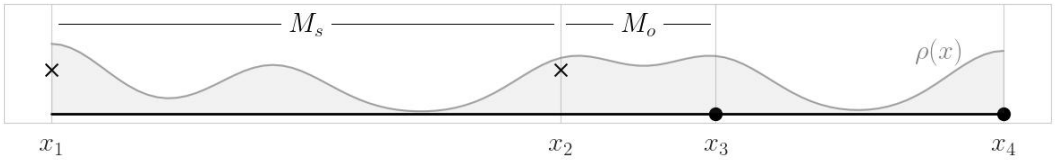


Figure 10. *Visualization of the 1D continuum example setup. The density $\rho(x)$ is plotted in gray, while the labeled points x_1, x_2, x_3, x_4 are plotted where the corresponding label is denoted by $\times$ or a solid dot. M_s marks the length between two similarly labeled points, while M_o marks the length between two oppositely labeled points.*

Solving the PWLL-τ equation[6] (4.1) on Ω can be broken into subproblems defined on the intervals $(x_1, x_2), \ldots, (x_{\ell-1}, x_\ell) \subset \mathbb{R}$, with boundary conditions determined by the corresponding labels of the endpoints x_i. There are two separate kinds of subproblems to be solved, as determined by these boundary conditions; namely, (1) the *oppositely labeled problem* (when $y_i \neq y_{i+1}$) and (2) the *similarly labeled problem* (when $y_i = y_{i+1}$). Recall from (4.1) that the boundary conditions for solutions u_1, u_2 at the labeled points correspond to the entries of the one-hot encoding of the labels $y_i \in \{1, 2\}$. For example, if $y_i = 2$ then $u_1(x_i) = 0, u_2(x_i) = 1$ will be the respective Dirichlet boundary conditions at the labeled point $x_i \in \mathcal{L}$.

Given the current labeled data, the active learning process selects a new query point $x^* = \operatorname{argmin}_{x \in \Omega} \mathcal{A}(x)$ via the minimum norm acquisition function (4.2). We can quantify the explorative behavior of our acquisition function (4.2) by comparing the minimizers of $\mathcal{A}(x)$ in (i) an interval of length M_o between *oppositely labeled points* and (ii) an interval of length M_s between *similarly labeled points*. In this simple one-dimensional problem, we may characterize "explorative" query points as residing in relatively *large* intervals between labeled points, regardless of the labels of the endpoints. Conversely, we characterize "exploitative" query points as residing between *oppositely labeled points that are close together*. In Figure 10, exploration would correspond to sampling in (x_1, x_2) or (x_3, x_4), while exploitation would correspond to sampling in (x_2, x_3).

[6] Without the reweighting (4.4) due to the simple geometry in one dimension.

The acquisition function (4.2) is directly a function of the magnitudes of the solutions to (4.1) with the corresponding boundary conditions; the decay of these solutions depends on the value of $\tau > 0$ (Figure 1). As such, we identify how τ must be chosen in order to produce small acquisition function values between similarly labeled points in relatively large regions as compared to large values in relatively small regions between oppositely labeled points.

In order to rigorously quantify the choice of $\tau > 0$, we give the mild assumptions that the density $\rho(x)$ (i) is sufficiently smooth, (ii) is *symmetric about the midpoint of the interval* between similarly labeled points, and (iii) obeys *a bounded derivative condition at the ends of the interval* between oppositely labeled points. Under these mild assumptions, we give the following simplified guarantee on exploration, which we prove rigorously in Section **??**.

Proposition 4.1 (Simplified version of Proposition **??**). *Suppose that the density $\rho(x)$ satisfies the above assumptions. Let the interval length M_o be relatively small compared to M_s; i.e., $M_o = \beta M_s$ for some $\beta \leq \frac{1}{4}$. Then we are ensured that*

$$\min_x \ \mathcal{A}_s(x) < \min_x \ \mathcal{A}_o(x)$$

as long as $\tau > 0$ and M_s jointly satisfy the following inequality

$$M_s^2 \left(C_0(\rho_s)\sqrt{\tau} - C_1(\rho_o)\beta^2\tau \right) \geq 8 \ln 2, \tag{4.3}$$

where $C_0(\rho_s)$ and $C_1(\rho_o)$ are constants that depend on the density ρ on the similarly and oppositely labeled intervals, respectively denoted ρ_s and ρ_o.

As long as the similarly labeled region has significantly large regions where the density $\rho(x)$ is sufficiently small compared to the oppositely labeled region, then we can be assured that choosing $\tau > 0$ large enough will result in query points between similarly labeled points that are relatively far from each other (as quantified by $\beta > 0$). We refer the reader to **??** in the Supplemental Material for further discussion of this result.

4.2. Exploration bounds in arbitrary dimensions. In this section, we show how larger values for τ lead to explorative behaviour in higher dimensional problems. In particular, we show that the acquisition function $\mathcal{A}(x)$ is small on unexplored clusters, and large on sufficiently well-explored clusters. This ensures that adequate exploration occurs before exploitation.

Let us remark that the reweighting term γ must be sufficiently singular near the labels $\mathcal{L}$ to ensure that (4.1) is well-posed. We recall from [17] that we require that γ has the form

$$\gamma(x) = 1 + \operatorname{dist}(x, \mathcal{L})^{-\alpha}, \tag{4.4}$$

where $\alpha > d - 2$. In practice, we choose γ as the solution of the graph Poisson equation (2.3) introduced earlier. To make the analysis in this section tractable, we assume here that γ satisfies (4.4), as was assumed in [17]. We emphasize here that without the singular reweighting γ, the equation (4.1) is ill-posed when the label set $\mathcal{L}$ is finite, and as such, there is no continuum version of active learning for us to study.

For an open set $A \subset \mathbb{R}^d$ and $r > 0$ we define the nonlocal boundary $\partial_r A$ as

$$\partial_r A = \overline{(A + B_r)} \setminus A.$$

The nonlocal boundary is essentially a tube of radius r surrounding the set A. The usual boundary is obtained by taking $r = 0$, so $\partial A = \partial_0 A$.

Our first result concerns upper bounds on the acquisition function in an unexplored cluster.

Theorem 4.2. *Let $\tau \geq 0$, $s, R > 0$ and $\mathcal{D} \subset \Omega$ with $\partial_{2s}\mathcal{D} \subset \Omega$ and $\mathcal{L} \cap (\mathcal{D} + B_{R+2s}) = \varnothing$. Let $\delta = \max_{\partial_{2s}\mathcal{D}} \rho$. Then the following hold.*

(i) *If*

$$\sqrt{\frac{\tau}{\delta}} \geq 3\left(\frac{d}{s} + 2\|\nabla \log \rho\|_{L^\infty(\partial_s \mathcal{D})}\right)(1 + R^{-\alpha}) + 3R^{-\alpha-1} \tag{4.5}$$

then we have that

$$\sup_{\mathcal{D}} \mathcal{A} \leq \sqrt{C} \exp\left(-\frac{s}{4}\sqrt{\frac{\tau}{\delta}}\right). \tag{4.6}$$

(ii) *Suppose that $B_r(x) \subset \mathcal{D}$ and let $M = \sup_{B_r(x)} \rho$. If (4.5) holds and*

$$\sqrt{\frac{\tau}{M}} \geq 3\left(\frac{d}{r} + 2\|\nabla \log \rho\|_{L^\infty(B_r(x_0))}\right)(1 + R^{-\alpha}) + 3R^{-\alpha-1} \tag{4.7}$$

then we have that

$$\sup_{B_{\frac{r}{2}}(x)} \mathcal{A} \leq \sqrt{C} \exp\left(-\frac{1}{4}\left(s\sqrt{\frac{\tau}{\delta}} + r\sqrt{\frac{\tau}{M}}\right)\right). \tag{4.8}$$

Remark 4.3. Theorem 4.2(i) shows that the acquisition function $\mathcal{A}$ is exponentially small on an unexplored cluster $\mathcal{D}$ provided there is a thin surrounding set $\partial_s \mathcal{D}$ of the cluster on which the density is small (less than δ), relatively smooth (so $\nabla \log \rho$ is not too large), and relatively far away from other labeled data points (so that R is not too small). All of these smallness assumptions are relative to the size of the ratio τ/δ as expressed in (4.5). In particular, regardless of the size of the right-hand side in (4.5), the condition can always be satisfied if the ratio τ/δ is sufficiently large, so we can view (4.5) as a condition on how small δ must be (i.e., how isolated $\mathcal{D}$ must be from other clusters).

Theorem 4.2(ii) improves the result in part (i) when $\mathcal{D}$ is a large cluster, in the sense that a large ball $B_r(x)$ fits inside $\mathcal{D}$. In this case, we expect the density ρ to be large within the cluster, so M will possibly be large relative to τ, and the estimate (4.8) is only a significant improvement to (4.6) when r is also large, that is, the cluster $\mathcal{D}$ has a large diameter. Hence, we can view (4.7) as a condition on how large r and R must be, and how small $\|\nabla \log \rho\|_{L^\infty(B_r(x_0))}$ must be, in order to obtain further exponential decay of the acquisition function within $\mathcal{D}$. In particular, regardless of how small τ/M is, the condition (4.7) will hold for large enough r, R and small enough $\|\nabla \log \rho\|_{L^\infty(B_r(x_0))}$ (i.e., the density is roughly constant within a ball in the cluster). We also mention that in 4.2(ii) we do not require δ to be small; that is, we do not require $\mathcal{D}$ to be a cluster that is separated from the rest of the dataset in order to have exponential decay of the acquisition function. Thus, 4.2(ii) applies to datasets that do not admit a clusterability structure.

However, we caution the reader that Theorem 4.2(ii) does not imply that our method will always choose the largest unexplored cluster to label next. The estimates in the theorem are upper bounds; they are quite likely loose and corresponding lower bounds (on unexplored clusters) do not exist. The question of which cluster will be sampled next depends also on the geometric arrangement of the clusters relative to the existing labeled data points, which is not addressed by the theorem. That is, a small cluster located very far away from existing labeled data points may be sampled prior to a large cluster that is much closer to the labeled data. In many situations, this is a completely reasonable action to take, and we would argue that it is not always desirable to choose the largest unexplored cluster next.

To ensure that new clusters are explored, we also need to lower bound the acquisition function near the existing labeled set. To do this, we need to introduce a model for the clusterability of the dataset. Let $\Omega_1, \Omega_2, \dots, \Omega_K \subset \Omega$ be disjoint sets representing each of the K clusters in the dataset. There are generally more clusters than classes ($K \geq C$) and often $K \gg C$. We assume there is a positive separation between clusters, measured by the quantity

$$\mathcal{S} := \min_{i \neq j} \operatorname{dist}(\Omega_i, \Omega_j). \tag{4.9}$$

The definition of $\mathcal{S}$ implies that $(\Omega_i + B_{\mathcal{S}}) \cap \Omega_j = \varnothing$ for all $i \neq j$. We define the union of the clusters as $\Omega' = \cup_{i=1}^K \Omega_i$. We note that we do not have $\Omega' = \Omega$, and it is important that there is room in the background $\Omega \setminus \Omega'$, which provides a separation between clusters. The background $\Omega \setminus \Omega'$ may have low density (though we do not assume this below), and can consist of outliers or data points that have characteristics of multiple classes and may be hard to classify.

Theorem 4.4. *Let $\tau \geq 0$ and $\alpha > d - 2$. Let $r > 0$ be small enough so that $r \leq \frac{1}{4}\mathcal{S}$,*

$$\tau r^d \leq \frac{1}{2^d 9}(\alpha + 2 - d)^2 \inf_{\Omega'} \rho, \tag{4.10}$$

and

$$4\|\nabla \log \rho\|_{L^\infty(\Omega')}(1 + 2^\alpha r^\alpha) r + \alpha 2^\alpha r^\alpha \leq \frac{1}{4}(\alpha + 2 - d). \tag{4.11}$$

Assume that $\mathcal{L} + B_{2r} \subset \Omega'$. Then we have

$$\inf_{\mathcal{L} + B_r} \mathcal{A} \geq 1 - 2^{-\frac{1}{2}(\alpha + 2 - d)}. \tag{4.12}$$

We now combine Theorems 4.2 and 4.4 to obtain a sample complexity result for the exploration performance of our algorithm. We need to introduce some notation for this. For $\mathcal{D} \subset \Omega$ we define $\mathcal{D}_\varepsilon = \{x \in \mathcal{D} : B_\varepsilon(x) \subset \mathcal{D}\}$. We define an *$\varepsilon$-packing* of Ω_i as a disjoint union of ε-balls that are centered at points in Ω_i. The *ε-packing number* of Ω_i is defined as

$$M(\Omega_i, \varepsilon) = \max\{m : \text{there exists an } \varepsilon\text{-packing of } \Omega_i \text{ with } m \text{ balls.}\}.$$

We can now state our result on sample complexity.

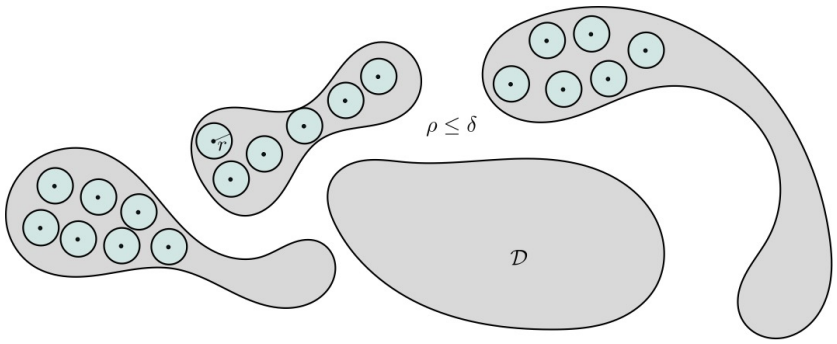


Figure 11. *Illustration of the implications of Theorems 4.2 and 4.4, and the discussion in Remark 4.6. The gray regions are the 4 clusters of high density in the dataset, and the density is small $\rho \le \delta$ between clusters. The current labeled set are the points at the centers of the blue balls. Theorems 4.2 and 4.4 guarantee that the next labeled point cannot lie in any of the blue balls, which correspond to the dilated label set $\mathcal{L} + B_r$. Once the dilated labels cover the existing clusters, the algorithm is guaranteed to select a point from the unexplored cluster $\mathcal{D}$. The number of labeled points selected from a given cluster during exploration is bounded by its $\frac{r}{2}$-packing number, as explained in Remark 4.6.*

Theorem 4.5 (Sample Complexity). *Let $\alpha > d - 2$. Let $R = s = \frac{1}{4}\mathcal{S}$ in Theorem 4.2 and choose τ to ensure (4.5) holds with $\mathcal{D} = \Omega_i + B_s$ for every i, where*

$$\delta = \max_{\Omega \setminus (\Omega' + B_s)} \rho.$$

Choose $r > 0$ to satisfy the conditions in Theorem 4.4 and assume that

$$\sqrt{C} \exp\left(-\frac{\mathcal{S}}{16}\sqrt{\frac{\tau}{\delta}}\right) \le 1 - 2^{-\frac{1}{2}(\alpha+2-d)}. \tag{4.13}$$

If the next active learning point is chosen sequentially to minimize the acquisition function $\mathcal{A}$ over Ω'_r then the algorithm will choose at most $M(\Omega_i, \frac{r}{2})$ points from Ω_i before all other clusters in $\Omega' \setminus \Omega_i$ have been sampled at least once. In particular, the algorithm will sample from all clusters within the first $\sum_{i=1}^{K} M(\Omega_i, \frac{r}{2})$ samples.

Remark 4.6. Theorem 4.5 shows that the number of samples required to explore all clusters in the dataset is $O(K)$, where the constant depends on the geometry and clusterability properties of the dataset (i.e., the packing numbers of the clusters). Thus, the method is very efficient at exploring the dataset in the early stages of active learning when τ is large. We can compare this to random sampling, which is also guaranteed to eventually explore all clusters, but takes in expectation $O(K \log(K))$ samples to do so (i.e., the coupon collector problem). Thus, our method improves on random sampling by a $\log(K)$ factor. We can see this improvement over random sampling and other existing methods in our experimental results. For example, on the **MNIST** dataset, Figure 7 shows that our algorithm explores all clusters after only 10 active learning iterations, at which point random sampling has explored less than 80% of the clusters (random sampling does not get close to full exploration until 30 iterations). On **FASHIONMNIST** (see Figure 8) we explore all clusters by 20 iterations,

at which point random sampling has explored around 90% of clusters. Similar results are observed on the **EMNIST** dataset in Figure 9.

In the idealized case of a dataset comprised of disconnected clusters (i.e., the background density between clusters Ω_i is $\delta = 0$), our theory would imply that K samples would ensure the exploration of all clusters, regardless of the size of clusters. While this would prove to be a further improvement over naive random sampling in the case of very disparately sized clusters, we note that the exploration of clusters would not be unique to our proposed acquisition function in practice since the identification of clusters would be immediate from the connected components of the similarity graph.

Remark 4.7. We note that the choices of the parameters r, s, R and τ are all dependent on the domain, the clusterability assumption, and the density, but are independent of the choices of labeled points $\mathcal{L}_i$. We also mention that there is an assumption made in Theorem 4.5 that there are no labeled points selected near the background region $\Omega \setminus \Omega'$. Indeed, if such outlying data points are selected as labeled points, then our results do not hold. In practice, one can perform sampling proportional to a density estimation, or simply remove outliers, to avoid such an issue. We discuss how this can be done in Supplemental Material Section **??**, and we have performed experiments with this. We have found that our experimental results are similar with and without outlier removal. We accordingly see this as an extra step that one has the option of performing in practice in order to maximally align the algorithm with the theory, but we do not see it as a necessary step in practice.

Remark 4.8. We also mention that there are certain features of the PWLL model that are used in the theoretical results in this section; namely, the continuum limit PDE is well-posed with arbitrarily few labels, and it satisfies a maximum (or comparison) principle, which is the main tool in our proofs. The p-Laplace models (see [10, 28, 64]) also satisfy these conditions when $p > d$ where d is the intrinsic dimension of the ambient space (or underlying manifold), and we fully expect that some results analogous to those in this section would hold for the p-Laplacian. We leave such investigations to future work and simply note here that solving the p-Laplace equation on a graph is far more computationally complex than the linear equation that constitutes PWLL. Thus, p-Laplace learning is not ideal for use in active learning, where the model is constantly re-evaluated throughout the active learning process.

Remark 4.9. In similar fashion to Remark 4.8, the theoretical tools we utilize for proving exploration guarantees of the PWLL model do not readily apply to methods like VOpt [41] and ΣOpt [50] due to the lack of a well-defined continuum limit of the (non-reweighted) Laplace learning model. The theoretical work for those acquisition functions presented in [50] focused on guarantees of greedy optimization of submodular set functions over finite sets, which can reasonably be assumed to imply exploration of the dataset in practice. Howevever, as of the writing of this paper, the authors are not aware of explicit theoretical guarantees for exploration in active learning similar to our work or previous works in active learning [18, 44, 55]. Furthermore, our maximum principle arguments are tailored to the simple and efficient-to-compute acquisition function (Unc. Norm) that is a function of the semi-supervised classifier at each active learning iteration; in contrast, the VOpt and ΣOpt acquisition functions are computationally expensive and are derived from the underlying differential operator, not the

semi-supervised classifier, at each iteration. Potential future work could investigate how a reweighting of the differential operator (as done in the PWLL model) may allow for exploration guarantees for the VOpt and ΣOpt acquisition functions.

5. Conclusion. We have demonstrated that uncertainty sampling is sufficient for exploration in graph-based active learning by using the norm of the output node function of the PWLL-τ model as an acquisition function. We provide rigorous mathematical guarantees on the explorative behavior of the proposed acquisition function. This is made possible by the well-posedness of the corresponding continuum limit PDE of the PWLL-τ model. Our analysis elucidates how the choice of hyperparamter $\tau > 0$ directly influences these guarantees; in the one dimensional case this effect is most clearly illustrated. In addition, we provide numerical experiments that further illustrate the effect of both our acquisition function and the hyperparameter τ on the sequence of active learning query points. Other numerical experiments confirm our theoretical guarantees and demonstrate favorable performance in terms of both accuracy and cluster exploration.

REFERENCES

[1] M.-F. Balcan, A. Beygelzimer, and J. Langford, *Agnostic active learning*, J. of Computer and System Sciences, 75 (2009), pp. 78–89, https://doi.org/10.1016/j.jcss.2008.07.003.

[2] M.-F. Balcan, A. Broder, and T. Zhang, *Margin based active learning*, in International Conference on Computational Learning Theory, vol. 4539, Springer Berlin Heidelberg, 2007, pp. 35–50, https://doi.org/10.1007/978-3-540-72927-3_5.

[3] M. Belkin, I. Matveeva, and P. Niyogi, *Regularization and semi-supervised learning on large graphs*, in Learning Theory: 17th Annual Conference on Learning Theory, COLT 2004, Banff, Canada, July 1-4, 2004. Proceedings 17, Springer, 2004, pp. 624–638.

[4] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural computation, 15 (2003), pp. 1373–1396.

[5] M. Belkin and P. Niyogi, *Semi-supervised learning on Riemannian manifolds*, Machine learning, 56 (2004), pp. 209–239.

[6] Y. Bengio, O. Delalleau, and N. Le Roux, *Label Propagation and Quadratic Criterion*, MIT Press, semi-supervised learning ed., January 2006, pp. 193–216, https://www.microsoft.com/en-us/research/publication/label-propagation-and-quadratic-criterion/.

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

[8] A. L. Bertozzi and E. Merkurjev, *Graph-based optimization approaches for machine learning, uncertainty quantification and networks*, in Handbook of Numerical Analysis, vol. 20, Elsevier, 2019, pp. 503–531.

[9] H. Cai, V. W. Zheng, and K. C.-C. Chang, *Active learning for graph embedding*, preprint arXiv, (2017), https://arxiv.org/abs/1705.05085.

[10] J. Calder, *The game theoretic p-Laplacian and semi-supervised learning with few labels*, Nonlinearity, 32 (2018), p. 301.

[11] J. Calder, *Consistency of Lipschitz learning with infinite unlabeled data and finite labeled data*, SIAM J. on Mathematics of Data Science, 1 (2019), pp. 780–812, https://doi.org/10.1137/18m1199241.

[12] J. Calder, *GraphLearning Python package*, Jan. 2022, https://doi.org/10.5281/zenodo.5850940.

[13] J. Calder, B. Cook, M. Thorpe, and D. Slepčev, *Poisson learning: Graph-based semi-supervised learning at very low label rates*, in Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, Nov. 2020, pp. 1306–1316.

[14] J. Calder, B. Cook, M. Thorpe, D. Slepčev, Y. Zhang, and S. Ke, *Graph-based semi-supervised learning with Poisson equations*, In preparation, (2022).

[15] J. Calder and N. García Trillos, *Improved spectral convergence rates for graph Laplacians on ε-graphs and k-NN graphs*, Applied and Computational Harmonic Analysis, 60 (2022), pp. 123–175, https://doi.org/10.1016/j.acha.2022.02.004.
[16] J. Calder, D. Slepčev, and M. Thorpe, *Rates of convergence for Laplacian semi-supervised learning with low labeling rates*, preprint arXiv, (2020), http://arxiv.org/abs/2006.02765.
[17] J. Calder and D. Slepčev, *Properly-weighted graph Laplacian for semi-supervised learning*, Applied Mathematics and Optimization: Special Issue on Optimization in Data Science, 82 (2020), pp. 1111–1159, https://doi.org/10.1007/s00245-019-09637-3.
[18] A. Cloninger and H. N. Mhaskar, *Cautious active clustering*, Applied and Computational Harmonic Analysis, 54 (2021), pp. 44–74, https://doi.org/10.1016/j.acha.2021.02.002.
[19] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, *EMNIST: An extension of MNIST to handwritten letters*, 2017, http://arxiv.org/abs/1702.05373.
[20] R. R. Coifman and S. Lafon, *Diffusion maps*, Applied and computational harmonic analysis, 21 (2006), pp. 5–30.
[21] G. Dasarathy, R. Nowak, and X. Zhu, *S2: An efficient graph based active learning algorithm with application to nonparametric classification*, in Proceedings of The 28th Conference on Learning Theory, P. Grünwald, E. Hazan, and S. Kale, eds., vol. 40 of Proceedings of Machine Learning Research, Paris, France, 2015, pp. 503–522.
[22] S. Dasgupta, *Coarse sample complexity bounds for active learning*, in Advances in Neural Information Processing Systems, vol. 18, MIT Press, 2006, pp. 235–242.
[23] S. Dasgupta, *Two faces of active learning*, Theoretical Computer Science, 412 (2011), pp. 1767–1781, https://doi.org/10.1016/j.tcs.2010.12.054, https://doi.org/10.1016/j.tcs.2010.12.054.
[24] S. Dasgupta and D. Hsu, *Hierarchical sampling for active learning*, in Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, July 2008, Association for Computing Machinery, pp. 208–215, https://doi.org/10.1145/1390156.1390183.
[25] D. L. Donoho and C. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data*, Proceedings of the National Academy of Sciences, 100 (2003), pp. 5591–5596.
[26] D. Dua and C. Graff, *UCI machine learning repository*, 2017, http://archive.ics.uci.edu/ml.
[27] M. M. Dunlop, D. Slepčev, A. M. Stuart, and M. Thorpe, *Large data and zero noise limits of graph-based semi-supervised learning algorithms*, Applied and Computational Harmonic Analysis, 49 (2020), pp. 655–697.
[28] A. El Alaoui, X. Cheng, A. Ramdas, M. J. Wainwright, and M. I. Jordan, *Asymptotic behavior ofℓ_p-based Laplacian regularization in semi-supervised learning*, in Conference on Learning Theory, PMLR, 2016, pp. 879–906.
[29] M. Flores, J. Calder, and G. Lerman, *Analysis and algorithms for ℓ_p-based semi-supervised learning on graphs*, Applied and Computational Harmonic Analysis, 60 (2022), pp. 77–122.
[30] Y. Gal, R. Islam, and Z. Ghahramani, *Deep Bayesian active learning with image data*, in Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 2017, Journal of Machine Learning Research, pp. 1183–1192.
[31] L. Gao, H. Yang, C. Zhou, J. Wu, S. Pan, and Y. Hu, *Active discriminative network representation learning*, in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 2142–2148, https://doi.org/10.24963/ijcai.2018/296.
[32] N. García Trillos, M. Gerlach, M. Hein, and D. Slepčev, *Error estimates for spectral convergence of the graph Laplacian on random geometric graphs toward the Laplace–Beltrami operator*, Foundations of Computational Mathematics, 20 (2020), pp. 827–887.
[33] J. Ham, D. Lee, and L. Saul, *Semisupervised alignment of manifolds*, in International Workshop on Artificial Intelligence and Statistics, PMLR, 2005, pp. 120–127.
[34] S. Hanneke, *A bound on the label complexity of agnostic active learning*, in Proceedings of the 24th International Conference on Machine Learning, New York, NY, USA, 2007, Association for Computing Machinery, pp. 353–360, https://doi.org/10.1145/1273496.1273541.
[35] S. Hanneke, *Theory of disagreement-based active learning*, Foundations and Trends® in Machine Learning, 7 (2014), pp. 131–309, https://doi.org/10.1561/2200000037.
[36] S. Hanneke and L. Yang, *Minimax analysis of active learning*, Journal of Machine Learning Research,

16 (2015), pp. 3487–3602, http://jmlr.org/papers/v16/hanneke15a.html.
[37] Y. He, W. Liang, D. Zhao, H.-Y. Zhou, W. Ge, Y. Yu, and W. Zhang, *Attribute surrogates learning and spectral tokens pooling in transformers for few-shot learning*, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 9109–9119, https://doi.org/10.1109/CVPR52688.2022.00891.
[38] M. Hein, J.-Y. Audibert, and U. v. Luxburg, *Graph Laplacians and their convergence on random neighborhood graphs.*, Journal of Machine Learning Research, 8 (2007).
[39] M. Hein, J.-Y. Audibert, and U. Von Luxburg, *From graphs to manifolds-weak and strong pointwise consistency of graph laplacians.*, in COLT, vol. 3559, Springer, 2005, pp. 470–485.
[40] S. Hu, Z. Xiong, M. Qu, X. Yuan, M.-A. Côté, Z. Liu, and J. Tang, *Graph policy network for transferable active learning on graphs*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 10174–10185.
[41] M. Ji and J. Han, *A variance minimization criterion to active learning on graphs*, in Artificial Intelligence and Statistics, Mar. 2012, pp. 556–564.
[42] H. Jiang and M. Gupta, *Minimum-margin active learning*, preprint arXiv, (2019), https://arxiv.org/abs/1906.00025.
[43] K.-S. Jun and R. Nowak, *Graph-based active learning: A new look at expected error minimization*, in 2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Dec. 2016, pp. 1325–1329, https://doi.org/10.1109/GlobalSIP.2016.7906056.
[44] M. Karzand and R. D. Nowak, *Maximin active learning in overparameterized model classes*, IEEE Journal on Selected Areas in Information Theory, 1 (2020), pp. 167–177, https://doi.org/10.1109/JSAIT.2020.2991518.
[45] D. P. Kingma and M. Welling, *Auto-encoding variational Bayes*, preprint arXiv, (2013), https://arxiv.org/abs/1312.6114.
[46] D. P. Kingma and M. Welling, *An introduction to variational autoencoders*, Foundations and Trends® in Machine Learning, 12 (2019), pp. 307–392, https://doi.org/10.1561/2200000056. Publisher: Now Publishers, Inc.
[47] D. Kushnir and L. Venturi, *Diffusion-based deep active learning*, preprint arXiv, (2020), https://arxiv.org/abs/2003.10339.
[48] Y. LeCun and C. Cortes, *MNIST handwritten digit database*, (2010), http://yann.lecun.com/exdb/mnist/.
[49] W.-Y. Lee, L.-C. Hsieh, G.-L. Wu, and W. Hsu, *Graph-based semi-supervised learning with multi-modality propagation for large-scale image datasets*, Journal of visual communication and image representation, 24 (2013), pp. 295–302.
[50] Y. Ma, R. Garnett, and J. Schneider, *Σ-optimality for active learning on Gaussian random fields*, in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., Curran Associates, Inc., 2013, pp. 2751–2759.
[51] K. Miller and A. L. Bertozzi, *Model-change active learning in graph-based semi-supervised learning*, preprint arXiv, (2021), http://arxiv.org/abs/2110.07739.
[52] K. Miller, H. Li, and A. L. Bertozzi, *Efficient graph-based active learning with probit likelihood via Gaussian approximations*, in ICML Workshop on Experimental Design and Active Learning, July 2020, http://arxiv.org/abs/2007.11126.
[53] K. Miller, J. Mauro, J. Setiadi, X. Baca, Z. Shi, J. Calder, and A. Bertozzi, *Graph-based active learning for semi-supervised classification of sar data*, in Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE) 2022 Conference on Defense + Commercial Sensing, SPIE, 2022.
[54] B. Mirzasoleiman, *Big Data Summarization Using Submodular Functions*, PhD thesis, ETH Zurich, 2017.
[55] J. M. Murphy and M. Maggioni, *Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion*, IEEE Transactions on Geoscience and Remote Sensing, 57 (2019), pp. 1829–1845, https://doi.org/10.1109/TGRS.2018.2869723.
[56] B. Nadler, N. Srebro, and X. Zhou, *Semi-supervised learning with the graph Laplacian: The limit of infinite unlabelled data*, in Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS'09, Red Hook, NY, USA, 2009, Curran Associates Inc., p. 1330–1338.

[57] Y.-L. Qiao, C. X. Shi, C. Wang, H. Li, M. Haberland, X. Luo, A. M. Stuart, and A. L. Bertozzi, *Uncertainty quantification for semi-supervised multi-class classification in image processing and ego-motion analysis of body-worn videos*, Image Processing: Algorithms and Systems, (2019), https://doi.org/10.2352/issn.2470-1173.2019.11.ipas-264.

[58] P. Sellars, A. I. Aviles-Rivero, and C.-B. Schönlieb, *Laplacenet: A hybrid graph-energy neural network for deep semisupervised classification*, IEEE Transactions on Neural Networks and Learning Systems, (2022), pp. 1–13, https://doi.org/10.1109/TNNLS.2022.3203315.

[59] O. Sener and S. Savarese, *Active learning for convolutional neural networks: A core-set approach*, preprint arXiv, (2018), http://arxiv.org/abs/1708.00489.

[60] B. Settles, *Active Learning*, vol. 6, Morgan & Claypool Publishers LLC, June 2012, https://doi.org/10.2200/s00429ed1v01y201207aim018.

[61] Z. Shi, S. Osher, and W. Zhu, *Weighted nonlocal Laplacian on interpolation from sparse data*, Journal of Scientific Computing, 73 (2017), pp. 1164–1177.

[62] C. Shui, F. Zhou, C. Gagné, and B. Wang, *Deep active learning: Unified and principled method for query and training*, preprint arXiv, (2020), http://arxiv.org/abs/1911.09162.

[63] O. Siméoni, M. Budnik, Y. Avrithis, and G. Gravier, *Rethinking deep active learning: Using unlabeled data at model training*, in The 25th International Conference on Pattern Recognition (ICPR), 2021, https://doi.org/10.1109/ICPR48806.2021.9412716.

[64] D. Slepcev and M. Thorpe, *Analysis of p-Laplacian regularization in semisupervised learning*, SIAM Journal on Mathematical Analysis, 51 (2019), pp. 2085–2120.

[65] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, *FixMatch: Simplifying semi-supervised learning with consistency and confidence*, in Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., vol. 33, Curran Associates, Inc., 2020, pp. 596–608.

[66] S. Tong and D. Koller, *Support vector machine active learning with applications to text classification*, Journal of Machine Learning Research, 2 (2001), pp. 45–66.

[67] S. Vahidian, B. Mirzasoleiman, and A. Cloninger, *Coresets for estimating means and mean square error with limited greedy samples*, in Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), Proceedings of Machine Learning Research, Aug. 2020, pp. 350–359.

[68] U. von Luxburg, *A tutorial on spectral clustering*, Statistics and Computing, 17 (2007), pp. 395–416, https://doi.org/10.1007/s11222-007-9033-z, https://doi.org/10.1007/s11222-007-9033-z (accessed 2020-10-28).

[69] B. Wang, Z. Tu, and J. K. Tsotsos, *Dynamic label propagation for semi-supervised multi-class multi-label classification*, in Proceedings of the IEEE international conference on computer vision, 2013, pp. 425–432.

[70] M. Welling and T. N. Kipf, *Semi-supervised classification with graph convolutional networks*, in J. International Conference on Learning Representations (ICLR 2017), 2016.

[71] H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms*, 2017, http://arxiv.org/abs/1708.07747.

[72] N. Zhang, L. Li, X. Chen, S. Deng, Z. Bi, C. Tan, F. Huang, and H. Chen, *Differentiable prompt makes pre-trained language models better few-shot learners*, in International Conference on Learning Representations, 2022, https://openreview.net/forum?id=ek9a0qIafW.

[73] Y. Zhang, H. Tong, Y. Xia, Y. Zhu, Y. Chi, and L. Ying, *Batch active learning with graph neural networks via multi-agent deep reinforcement learning*, vol. 36, Jun. 2022, pp. 9118–9126, https://doi.org/10.1609/aaai.v36i8.20897.

[74] M. Zheng, S. You, L. Huang, F. Wang, C. Qian, and C. Xu, *SimMatch: Semi-supervised learning with similarity matching*, in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 06 2022, pp. 14451–14461, https://doi.org/10.1109/CVPR52688.2022.01407.

[75] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, *Graph neural networks: A review of methods and applications*, AI Open, 1 (2020), pp. 57–81, https://doi.org/https://doi.org/10.1016/j.aiopen.2021.01.001.

[76] X. Zhou and M. Belkin, *Semi-supervised learning by higher order regularization*, in Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 892–900.

[77] D. Zhu, Z. Li, X. Wang, B. Gong, and T. Yang, *A robust zero-sum game framework for pool-based active learning*, in The 22nd International Conference on Artificial Intelligence and Statistics, Apr. 2019, pp. 517–526.

[78] X. Zhu, Z. Ghahramani, and J. Lafferty, *Semi-supervised learning using Gaussian fields and harmonic functions*, in Proceedings of the 20th International Conference on International Conference on Machine Learning, Washington, DC, USA, Aug. 2003, AAAI Press, pp. 912–919.

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