MaxiMin Active Learning in Overparameterized Model Classes

Mina Karzand and Robert D. Nowak

***Abstract*—Generating labeled training datasets has become a major bottleneck in Machine Learning (ML) pipelines. Active ML aims to address this issue by designing learning algorithms that automatically and adaptively select the most informative examples for labeling so that human time is not wasted labeling irrelevant, redundant, or trivial examples. This paper proposes a new approach to active ML with nonparametric or overparameterized models such as kernel methods and neural networks. In the context of binary classification, the new approach is shown to possess a variety of desirable properties that allow active learning algorithms to automatically and efficiently identify decision boundaries and data clusters.**

***Index Terms*—Active learning, overparameterized learning, neural networks, reproducing Kernel Hilbert spaces, pool-based learning.**

I. Introduction

THE FIELD of Machine Learning (ML) has advanced considerably in recent years, but mostly in well-defined domains using huge amounts of human-labeled training data. Machines can recognize objects in images and translate text, but they must be trained with more images and text than a person can see in nearly a lifetime. The computational complexity of training has been offset by recent technological advances, but the cost of training data is measured in terms of the human effort in labeling data. People are not getting faster nor cheaper, so generating labeled training datasets has become a major bottleneck in ML pipelines. Active ML aims to address this issue by designing learning algorithms that automatically and adaptively select the most informative examples for labeling so that human time is not wasted labeling irrelevant, redundant, or trivial examples. This paper explores active ML with nonparametric or overparameterized models such as kernel methods and neural networks.

Deep neural networks (DNNs) have revolutionized machine learning applications, and theoreticians have struggled to explain their surpising properties. DNNs are highly overparameterized and often fit perfectly to data, yet remarkably the learned models generalize well to new data. A mathematical understanding of this phenomenom is beginning to emerge [1], [2], [3], [4], [5], [6], [7], [8]. This work suggests that among all the networks that could be fit to the training data, the learning algorithms used in fitting favor networks with smaller weights, providing a sort of implicit regularization. With this in mind, researchers have shown that shallow (but wide) networks and classical kernel methods fit to the data but regularized to have small weights (e.g., minimum norm fit to data) can generalize well [2], [8], [9], [10].

Despite the recent success and new understanding of these systems, it still is a fact that learning good neural network models can require an enormous number of labeled data. The cost of obtaining labels can be prohibitive in many applications. This has prompted researchers to investigate active ML for kernel methods and neural networks [11], [12], [13], [14], [15], [16]. None of this work, however, directly addresses overparameterized and interpolating regime, which is the focus in this paper. Active ML algorithms have access to a large but unlabeled dataset of examples and sequentially select the most "informative" examples for labeling [17], [18] . This can reduce the total number of labeled examples needed to learn an accurate model.

Broadly speaking, active ML algorithms adaptively select examples for labeling based on two general strategies [19]. The first is to select examples that rule-out as many (incompatible) classifiers as possible at each step. In effect, this leads to algorithms that tend to label examples near decision boundaries. The second strategy involves discovering cluster structure in unlabeled data and labeling representative examples from each cluster. We show that our new MaxiMin active learning approach automatically exploits both these strategies, as depicted in Figure 1.

This paper builds on a new framework for active learning in the overparameterized and interpolationg regime, focusing on kernel methods and two-layer neural networks in the binary classification setting. The approach, called *MaxiMin Active Learning*, is based on mininum norm interpolating models. Roughly speaking, at each step of the learning process the maximin criterion requests a label for the example that is most difficult to interpolate. A minimum norm interpolating model is constructed for each possible example and the one yielding the largest norm indicates which example to label next. The rationale for the maximin criterion is that labeling the most challenging examples first may eliminate the need to label many of the other examples.

Manuscript received October 16, 2019; revised April 23, 2020; accepted April 26, 2020. Date of publication April 30, 2020; date of current version June 8, 2020. This work was supported in part by the Air Force Machine Learning Center of Excellence under Grant FA9550-18-1-0166. Parts of this article were presented at the 57th Annual Allerton Conference on Communication, Control, and Computing, 2019. *(Corresponding author: Mina Karzand.)*

Mina Karzand is with the Wisconsin Institute of Discovery, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: karzand@wisc.edu).

Robert D. Nowak is with the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: rdnowak@wisc.edu).

Digital Object Identifier 10.1109/JSAIT.2020.2991518

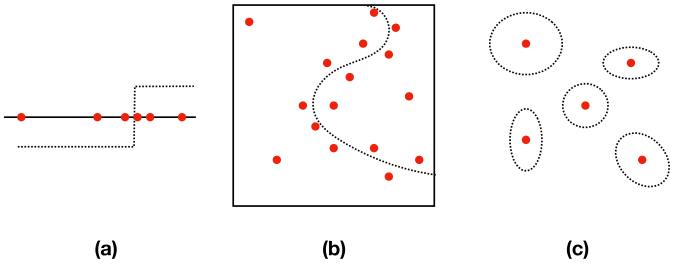

Fig. 1. MaxiMin Active Learning strategically selects examples for labeling (red points). (a) reduces to binary search in simple 1-d threshold problem setting; (b) labeling is focused near decision boundary in multidimensional setting; (c) automatically discovers clusters and labels representative examples from each.

The maximin selection criterion is studied through experiments and mathematical analysis. We prove that the criterion has a number of desirable properties:

- It tends to label examples near the current (estimated) decision boundary and close to oppositely labeled examples, allowing the active learning algorithm to focus on learning decision boundaries.
- It reduces to optimal bisection in the one-dimensional linear classifier setting.
- A data-based form of the criterion also provably discovers clusters and also automatically generates labeled coverings of the dataset.

Experimentally, we show that these properties generalize in several ways. For example, we find that in multiple dimensions the maximin criterion leads to a multidimensional bisection-like process that automatically finds a portion of the decision boundary and then locally explores to efficiently identify the complete boundary. We also show that MaxiMin Active Learning can learn hand-written digit classifiers with far fewer labeled examples than traditional passive learning based on labeling a randomly selected subset of examples.

II. A New Active Learning Criterion

At each iteration of the active learning algorithm, looking at the currently labeled set of samples, a new unlabeled point is selected to be labeled. The criterion we are proposing to pick the samples to be labeled is based on a 'maximin' operator. We will describe the criterion in its most general form along with the intuition behind this choice of criterion. In the remainder of the paper, we will go through some theoretical results about the properties of variations of this criterion in various setups along with some additional descriptive numerical evaluations and simulations.

A. Nonparametric Pool-Based Active Learning

At each time step, the algorithm has access to a pool of labeled samples and a set of unlabaled samples. In other words, we have a partially labeled training set. Let $\mathcal{L} = \{(x_1, y_1), \ldots, (x_L, y_L)\}$ be the set of labeled examples so far. We assume $x_i \in \mathcal{X}$ where $\mathcal{X}$ is the input/feature space and binary valued labels $y_i \in \{-1, +1\}$. Let $\mathcal{U} \subseteq \mathcal{X}$ be the set of unlabeled samples.

In the interpolating regime, the goal is to correctly label all the points in $\mathcal{U}$ so that the training error is zero: Passive learning generally requires labeling every point in $\mathcal{U}$. Active learning sequentially selects points in $\mathcal{U}$ for labeling with the aim of learning a correct classifier without necessarily labeling all of $\mathcal{U}$. Our setting can be viewed as an instance of *pool-based* active learning.

At each iteration, one unlabeled sample, $u^* \in \mathcal{U}$ is selected, labeled and added to the pool of labeled samples. The selection process is designed to pick the samples which are most *informative* upon being labeled. The proposed notion of score is the measure of informativeness of each sample $u \in \mathcal{U}$ at each time: the score of each unlabaled sample is computed, and the sample with the largest score is selected to be labeled.

$$u^* = \text{argmax}_{u \in \mathcal{U}} \textsf{score}(u). \tag{1}$$

If there are multiple maximizers, then one is selected uniformly at random. Note that for any unlabeled sample $u \in \mathcal{U}$, the value of $\textsf{score}(u)$ depends implicitly on the set of currently labeled points, $\mathcal{L}$. That is, information gained by labeling u depends on the current knowledge of the learner. To define our proposed notion of score, we define minimum norm interpolating function and introduce some notations next.

B. Minimum Norm Interpolating Function

Let $\mathcal{F}$ be a class of functions mapping $\mathcal{X}$ to $\mathbb{R}$, where $\mathcal{X}$ is the input/feature space,. We assume the class $\mathcal{F}$ is rich enough to interpolate the training data. For example, $\mathcal{F}$ could be a nonparametric infinite dimensional Reproducing Kernel Hilbert Space (RKHS) or an overparameterized neural network representation.

Given the set of labeled samples, $\mathcal{L}$, and a class of functions $\mathcal{F}$, let $f \in \mathcal{F}$ be the interpolating function such that $f(x_i) = y_i$ for all $(x_i, y_i) \in \mathcal{L}$. Note that there may be many functions that interpolate a discrete set of points such as $\mathcal{L}$. Among these, we choose f to be the minimum norm interpolator:

$$\begin{aligned} f(x) &\coloneqq \text{argmin}_{g \in \mathcal{F}} \ \ \|g\|_{\mathcal{F}} \\ \text{s.t.} \quad g(x_i) &= y_i, \ \text{ for all } (x_i, y_i) \in \mathcal{L}. \end{aligned} \tag{2}$$

Clearly, the definition of f depends on the set of currently labeled samples $\mathcal{L}$ and the function norm $\|\cdot\|_{\mathcal{F}}$, although we omit these dependencies for ease of notation. The choice of $\mathcal{F}$ and the norm $\|\cdot\|_{\mathcal{F}}$ is application dependent. In this paper, we focus on (1) function classes represented by an overparameterized neural network representation with the ℓ^2 norm of the weight vectors and (2) reproducing kernel Hilbert spaces with the corresponding Hilbert norm.

For unlabeled points $u \in \mathcal{U}$ and $\ell \in \{-1, +1\}$, define $f_\ell^u(x)$ is the minimum norm interpolating function based on current set of labeled samples $\mathcal{L}$ and the point $u \in \mathcal{U}$ with label ℓ:

$$\begin{aligned} f_\ell^u(x) &\coloneqq \text{argmin}_{g \in \mathcal{F}} \ \ \|g\|_{\mathcal{F}} \\ \text{s.t.} \quad g(x_i) &= y_i, \ \text{ for all } (x_i, y_i) \in \mathcal{L} \\ g(u) &= \ell. \end{aligned} \tag{3}$$

We use this definition in the next subsection to define the notion of score.

C. Definition of Proposed Notion of score

Roughly speaking, we want our selection criterion to prioritize labeling the most "informative" examples. Since the ultimate goal is to correctly label every example in $\mathcal{U}$, we design $\mathsf{score}(u)$ to measure the how hard it is to interpolate after adding u to the set of labeled points. The intuition is that attacking the most challenging points in the input space first may eliminate the need to label other 'easier' examples later.

Note that we need to compute $\mathsf{score}(u)$ without knowing the label of u. To do so, we come up with an estimate of label of u, denoted by $\ell(u) \in \{-1, +1\}$ and compute $\mathsf{score}(u)$ assuming that upon labeling, u will be labeled $\ell(u)$. We propose the following criterion for choosing $\ell(u)$:

$$\ell(u) \coloneqq \mathrm{argmin}_{\ell \in \{-1,+1\}} \|f^u_\ell(x)\|_{\mathcal{F}}. \tag{4}$$

Operating in the interpolating regime, we estimate the label of any unlabeled sample, u, to be the one that yields the minimum norm interpolant (i.e., the "smoother" of the two interpolants among the two possible functions $f^u_+(x)$ and $f^u_-(x)$).

Define

$$f^u(x) \coloneqq f^u_{\ell(u)}(x) \tag{5}$$

to be the interpolating function after adding the sample u with the label $\ell(u)$, defined in (4).

We propose two notions of score. For $u \in \mathcal{U}$, define

$$\mathsf{score}_{\mathcal{F}}(u) = \left\|f^u(x)\right\|_{\mathcal{F}} \tag{6}$$

$$\mathsf{score}_{\mathrm{D}}(u) = \left\|f^u(x) - f(x)\right\|_{\mathrm{D}} \tag{7}$$

where $\|\cdot\|_{\mathcal{F}}$ is the norm associated the function space $\mathcal{F}$. The function f is the minimum norm interpolator of the labeled examples in $\mathcal{L}$ (defined in (2)), and $f^u(x)$ is defined (5) as the minimum norm interpolator after adding u with the estimated label $\ell(u)$ to the set of labeled points. Also, define

$$\|g\|_{\mathrm{D}} = \int_{\mathcal{X}} |g(x)|^2 \, \mathrm{d}P_X(x), \tag{8}$$

where P_X is the distribution of x. In practice, P_X is the empirical distribution of $\mathcal{U}$. We refer to the (6) as the *function norm score* and (7) as the *data-based norm score*.[1]

The distinction between the two definitions of the score function is as follows. Scoring unlabeled points according to the definition $\mathsf{score}_{\mathcal{F}}$ priotorizes labeling the examples which result in minimum norm interpolating functions with largest norm. Since the norm of the function can be associated with its smoothness, roughly speaking, this means that this criterion picks the points which give the least smooth interpolating functions. However, $\mathsf{score}_{\mathcal{F}}$ is insensitive to the distribution of data. The data-based $\mathsf{score}_{\mathrm{D}}$, in contrast, is sensitive to the distribution of the data. Measuring the difference between the new interpolation f^u and the previous one makes this also sensitive to the structure of the function class.

With these definitions in place, we state the MaxiMin Active Learning criterion as follows. Given labeled data $\mathcal{L}$, the next example $u^* \in \mathcal{U}$ to label is selected according to

$$f^u = \arg \min_{f \in \{f^u_+, f^u_-\}} \|f\|_{\mathcal{F}} \ , \ \forall u \in \mathcal{U}$$

$$u^* = \arg \max_{u \in \mathcal{U}} \mathsf{score}(u)$$

with either $\mathsf{score}_{\mathcal{F}}$ or $\mathsf{score}_{\mathrm{D}}$.

III. MaxiMin Active Learning With Neural Networks

A. Overparameterized Neural Networks and Interpolation

Neural networks are often highly overparameterized and exactly fit to training data, yet remarkably the learned models generalize well to new data. A mathematical understanding of this phenomenom is beginning to emerge [1], [2], [3], [4], [5], [6], [7], [8]. This work suggests that among all the networks that could be fit to the training data, the learning algorithms used in training favor networks with smaller weights, providing a sort of implicit regularization. With this in mind, researchers have shown that even shallow networks and classical kernel methods fit to the data but regularized to have small weights (e.g., minimum norm fit to data) can generalize well [2], [8], [9], [10]. The functional mappings generated by wide, two-layer neural networks with Rectified Linear Unit (ReLU) activation functions were studied in [20]. It is shown that exactly fitting such networks to training data subject to minimizing the ℓ^2-norm of the network weights results in a linear spline interpolation. This result was extended to a broad class of interpolating splines by appropriate choices of activation functions [21]. Our analysis of the MaxiMin active learning with neural networks will leverage these connections.

B. Neural Network Regularization

It has been long understood that the size of neural network weights, rather than simply the number of weights/neurons, characterizes the complexity of neural networks [22]. Here we focus on two-layer neural networks with ReLU activation functions in the hidden layer. If $\boldsymbol{x} \in \mathbb{R}^d$ is input to the network, then the output is computed by the function

$$f_{\boldsymbol{w},\boldsymbol{b},c}(\boldsymbol{x}) = \sum_{n=1}^{N} v_n \, \sigma\left(\boldsymbol{u}_n^T \boldsymbol{x} + b_n\right) + c, \tag{9}$$

where $\sigma(\cdot) = \max\{0, \cdot\}$ is the ReLU activation, $\boldsymbol{w} \coloneqq \{v_n, \boldsymbol{u}_n\}_{n=1}^N$ are the "weights" of the network, and $\boldsymbol{b} \coloneqq \{b_n\}$ and c are constant "bias" terms. The "norm" of $f_{\boldsymbol{w},\boldsymbol{b},c}$ is defined as $\|f_{\boldsymbol{w},\boldsymbol{b},c}\| \coloneqq \|\boldsymbol{w}\|_2$, the ℓ_2-norm of the vector of network weights. We use the term norm in quotes because technically the weight norm does not correspond to a true norm on the function $f_{\boldsymbol{w},\boldsymbol{b},c}$ since, for example, constant functions $f_{\boldsymbol{w},\boldsymbol{b},c} = c$ have $\|\boldsymbol{w}\|_2 = 0$. From now on we will drop the subscripts and just write f for ease of notation. Let $\{(x_i, y_i)\}_{i=1}^M$ be a set of training data. The minimum "norm" neural network

[1]Operationally, to compute the data-based norm of any function, the algorithm uses the probability mass function of set of unlabeled points as a proxy for the input probability density function over the feature space $\mathcal{X}$. In particular, the algorithm approximates $\|g\|_{\mathrm{D}}$ by the average of the function over the set of unlabeled points: $\|g\|_{\mathrm{D}} \approx \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} |g(u)|^2$. High density of set of unlabeled points and some mild regularity conditions guarantee that this is a good approximation. Throughout the paper, we use (8) to prove theoretical statements and its approximation in the numerical simulations.

interpolation of these data is the solution to the optimization

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2 \text{ subject to } f(x_i) = y_i, \quad i = 1, \ldots, M.$$

A solution exists if the number of neurons N is sufficiently large (see [23, Th. 5.1]).

In Section V we explore the behavior of MaxiMin active learning through numerical experiments using both the function "norm" score and the data-based norm score. In all our experiments and theory, we assume the binary classification setting where $y_i = \pm 1$. Broadly speaking, we observe the following behaviors.

- With the function "norm" score the MaxiMin active learning algorithm tends to sample aggressively in the vicinity of the boundary, prefering to gather new labels between the closest oppositely labeled examples.
- The data-based norm score is sensitive to the distribution of the data. It strikes a balance between exploiting regions between oppositely labeled examples (as in the function-based case) and exploring regions further away from labeled examples. Thus we see evidence that the data-based norm can effectively seek out the decision boundary and explore data clusters.

These behaviors are supported by a formal analysis of MaxiMin active learning in one dimension, discussed next.

C. MaxiMin Active Learning in One-Dimension

Our analysis of MaxiMin active learning with neural networks will focus on the behavior in one-dimension. We show that MaxiMin active learning with a two-layer ReLU netwok recovers optimal bisection learning strategies. The following characterization of minimum "norm" neural network interpolation in one-dimension follows from [20], [21] (see [21, Th. 4.4 and Proposition 6.1]).

Theorem 1: Let $f : \mathbb{R} \to \mathbb{R}$ be a two-layer neural network with ReLU activation functions and N hidden nodes as in (9). Let $\{(x_i, y_i)\}_{i=1}^M$ be a set of training data. If $N \geq M$, then a solution to the optimization

$$\min_{\boldsymbol{w}} \|\boldsymbol{w}\|_2 \text{ subject to } f(x_i) = y_i, \quad i = 1, \ldots, M$$

is a minimal knot linear spline interpolation of the points $\{(x_i, y_i)\}_{i=1}^M$.

In our analysis, we exploit the equivalence between minimum "norm" neural networks and linear splines. Specifically, a solution to the optimization is an interpolating function that is linear between each pair of neighboring points. This ensures that given a pair of neighboring labeled points x_1 and x_2 and any unlabeled point $x_1 < u < x_2$, adding u to the set of labeled points can only potentially change the interpolating function between x_1 and x_2. To eliminate uncertainty in the boundary conditions of the interpolation, we assume that the neural network is initialized by labeling the leftmost and rightmost points in the dataset and forced to have a constant extension to the left and right of these points (this can be accomplished by adding two artificial points to the left and right with the same labels as the true endpoints).

The main message of our analysis is that MaxiMin active learning with two-layer ReLU networks recovers optimal bisection (binary search) in one-dimension. This is summarized by the next corollary which follows in a straightforward fashion from Theorems 2 and 3.

Corollary 1: Consider N points uniformly distributed in the interval $[0, 1]$ labeled according to a k-piecewise constant function f so that $y_i = f(x_i) \in \{-1, +1\}$, $i = 1, \ldots, N$, and length of the pieces are $\Theta(1/K)$. Then after labeling $O(k \log N)$ examples, the MaxiMin active learning with a two-layer ReLU network correctly labels all N examples (i.e., the training error is zero).

The corollary follows from the fact that the MaxiMin criteria (both function norm and data-based norm) selects the next example to label at the midpoint between neighboring and oppositely labeled examples (i.e., at a bisection point). This is characterized in the next two theorems. First we consider the function "norm" criterion. The proof of the following theorem appears in Appendix A.1.

Theorem 2: Let $\mathcal{L}$ be a set of labeled examples and let u be an unlabeled example. Let f_+^u be the minimum "norm" interpolator of $\mathcal{L} \cup (u, +1)$ and let f_-^u be the minimum "norm" interpolator of $\mathcal{L} \cup (u, -1)$. Define the score of an unlabeled example u as $\mathsf{score}_{\mathcal{F}}(u) = \min\{\|f_+^u\|, \|f_-^u\|\}$, where $\|f\| = \|\boldsymbol{w}\|_2$, the neural network weight norm. Then, the selection criterion based on $\mathsf{score}_{\mathcal{F}}$ has the following properties

1) Let x_1 and x_2 be two oppositely labeled neighboring points in $\mathcal{L}$, i.e., no other points between x_1 and x_2 have been labeled and $y_1 \neq y_2$. Then for all $x_1 < u < x_2$, $\mathsf{score}_{\mathcal{F}}(\frac{x_1+x_2}{2}) \geq \mathsf{score}_{\mathcal{F}}(u)$.
2) Let $x_1 < x_2$ and $x_3 < x_4$ be two pairs of oppositely labeled neighboring points (i.e., $y_1 \neq y_2$ and $y_3 \neq y_4$) such that $x_2 - x_1 \geq x_4 - x_3$. Then,
$$\mathsf{score}_{\mathcal{F}}\left(\frac{x_1 + x_2}{2}\right) \leq \mathsf{score}_{\mathcal{F}}\left(\frac{x_3 + x_4}{2}\right).$$
3) Let x_5 and x_6 be two identically labeled neighboring points in $\mathcal{L}$, i.e., $y_5 = y_6$. Then for all $x_5 < u < x_6$, the function $\mathsf{score}_{\mathcal{F}}(u)$ is constant.
4) For any pair of neighboring oppositely labeled points x_1 and x_2, any pair of neighboring identically labeled points x_5 and x_6, any $x_1 < u < x_2$ and any $x_5 < v < x_6$, we have
$$\mathsf{score}_{\mathcal{F}}(v) \leq \mathsf{score}_{\mathcal{F}}(u).$$

Now we turn to the data-based norm. Here we observe the effect of the data distribution on the bisection properties. The properties mirror those in Theorem 2 except in the case of the second property. The data-based norm criterion tends to sample in the *largest* (most data-massive) interval between oppositely labeled points, whereas the function-based norm criterion favors points in the *smallest* interval.

Theorem 3: Let the distribution $\mathbb{P}(X)$ be uniform over an interval. Let $\mathcal{L}$ be a set of labeled examples and let u be an unlabeled example. Let f_+^u be the minimum "norm" interpolator of $\mathcal{L} \cup (u, +1)$ and let f_-^u be the minimum "norm" interpolator of $\mathcal{L} \cup (u, -1)$ and let $f^u = \arg_{g \in \{f_+^u, f_-^u\}} \|g\|$ consistent with notations in (3) and (5). Then $\mathsf{score}_{\mathrm{D}}(u) = \int |f^u(x) - f(x)|^2 \, dP_X(x)$, where f is the minimum "norm"

interpolator based on the labeled data $\mathcal{L}$. Then, the selection criterion based on $\mathsf{score}_{\mathrm{D}}$ has the following properties.

1) Let x_1 and x_2 be two oppositely labeled neighboring points in $\mathcal{L}$, i.e., $y_1 \neq y_2$. Then for all $x_1 < u < x_2$ $\mathsf{score}_{\mathrm{D}}(\frac{x_1+x_2}{2}) \geq \mathsf{score}_{\mathrm{D}}(u)$.
2) Let $x_1 < x_2$ and $x_3 < x_4$ be two pairs of oppositely labeled neighboring labeled points (i.e., $y_1 \neq y_2$ and $y_3 \neq y_4$) such that $x_2 - x_1 \geq x_4 - x_3$. If the unlabeled points are uniformly distributed in each interval and the number of points is in (x_1, x_2) is less than the number in (x_4, x_3), then
$$\mathsf{score}_{\mathrm{D}}\left(\frac{x_1+x_2}{2}\right) \geq \mathsf{score}_{\mathrm{D}}\left(\frac{x_3+x_4}{2}\right).$$
3) Let x_5 and x_6 be two identically labeled neighboring points in $\mathcal{L}$, i.e., $y_5 = y_6$. Then for all $x_5 < v < x_6$, we have $\mathsf{score}_{\mathrm{D}}(v) = 0$.
4) For any pair of neighboring oppositely labeled points x_1 and x_2, any pair of neighboring identically labeled points x_5 and x_6, any $x_1 < u < x_2$ and any $x_5 < v < x_6$, we have
$$\mathsf{score}_{\mathrm{D}}(v) \leq \mathsf{score}_{\mathrm{D}}(u).$$

The proof appears in Appendix A.2.

IV. Interpolating Active Learners in an RKHS

In this section, we will focus on minimum norm interpolating functions in a Reproducing Kernel Hilbert Space (RKHS). We present theoretical properties for general RKHS settings, detailed analytical results in the one-dimensional setting, and numerical studies in multiple dimensions. Broadly speaking, we establish the following properties: the proposed score functions

- tend to select examples near the decision boundary of f, the current interpolator;
- the score is largest for unlabeled examples near the decision boundary *and* close to oppositely labeled examples, in effect searching for the boundary in the most likely region of the input space;
- in one dimension the interpolating active learner coincides with an optimal binary search procedure;
- using data-based function norms, rather than the RKHS norm, the interpolating active learner executes a tradeoff between sampling near the current decision boundary and sampling in regions far away from currently labeled examples, thus exploiting cluster structure in the data.

A. Kernel Methods

A Hilbert space $\mathcal{H}$ is associated with an inner product: $\langle f, g\rangle_{\mathcal{H}}$ for $f, g \in \mathcal{H}$. This induces a norm defined by $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$. A symmetric bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semidefinite if for all $n \geq 1$, and points $\{x_i\}_{i=1}^n$, the matrix $\mathbf{K}$ with element $\mathbf{K}_{i,j} = K(x_i, x_j)$ is positive semidefinite (PSD). These functions are called PSD kernel functions. A PSD kernel constructs a Hilbert space, $\mathcal{H}$ of functions on $f : \mathcal{X} \to \mathbb{R}$. For any $x \in \mathcal{X}$ and any $f \in \mathcal{H}$, the function $K(\cdot, x) \in \mathcal{H}$ and $\langle f, K(\cdot, x)\rangle_{\mathcal{H}} = f(x)$. Throughout this section, we assume $K(x, x) = 1$.

For the set of labeled samples $\mathcal{L} = \{(x_1, y_1), \ldots, (x_L, y_L)\}$ with $y_i \in \{-1, +1\}$, let the function $f(x)$ be decomposed as
$$f(x) = \sum_{i=1}^{L} \alpha_i k(x_i, x) \quad \text{with} \quad \alpha = \mathbf{K}^{-1}\mathbf{y}, \tag{10}$$
where $\mathbf{K} = [\mathbf{K}_{i,j}]_{i,j}$ is the L by L matrix such that $\mathbf{K}_{i,j} = k(x_i, x_j)$ and $\mathbf{y} = [y_1, \ldots, y_L]^T$. Using reproducible kernels implies that $f(x) \in \mathcal{H}$ for the a RKHS $\mathcal{H}$. Then, $f(x)$ defined above is the minimum Hilbert norm interpolating function defined in (2). Using the property $\langle K(x_i, \cdot), K(x_j, \cdot)\rangle = K(x_i, x_j)$, we have
$$\|f(x)\|_{\mathcal{H}}^2 = \alpha^T \mathbf{K} \alpha = \mathbf{y}^T \mathbf{K}^{-1} \mathbf{y}.$$

For $u \in \mathcal{U}$ and $\ell \in \{-1, +1\}$, the minimum norm interpolating unction $f_\ell^u(x)$, defined in (3) (based on currently labeled samples $\mathcal{L}$ and sample u with label ℓ) is derived similarly:
$$f_\ell^u(x) = \sum_{i=1}^{L} \widetilde{\alpha}_i k(x_i, x) + \widetilde{\alpha}_{L+1} k(u, x) \quad \text{with} \quad \widetilde{\alpha} = \widetilde{\mathbf{K}}_u^{-1} \widetilde{\mathbf{y}}_\ell, \tag{11}$$
where
$$\widetilde{\mathbf{K}}_u = \begin{bmatrix} \mathbf{K} & \mathbf{a}_u \\ \mathbf{a}_u^T & b \end{bmatrix}, \ \mathbf{a}_u = \begin{bmatrix} k(x_1, u) \\ \vdots \\ k(x_L, u) \end{bmatrix}, \ \widetilde{\mathbf{y}}_\ell = \begin{bmatrix} \mathbf{y} \\ \ell \end{bmatrix}, \quad \text{and } b = K(u, u). \tag{12}$$
Throughout this paper, we use kernel such that $K(x, x) = 1$ for all $x \in \mathcal{X}$.

B. Properties of General Kernels for Active Learning

We first show that using kernel based function spaces for interpolation, $\ell(u)$ defined in (4) coincides with the sign of value of current interpolator at u.

Proposition 1: For $u \in \mathcal{X}$ and $\ell \in \{-, +\}$, define $f(x)$ and $f_\ell^u(x)$ according to (2) and (3) in Section II. Then, $\ell(u)$ defined in (4) satisfies
$$\ell(u) = \begin{cases} +1 & \text{if } f(u) \geq 0 \\ -1 & \text{if } f(u) < 0. \end{cases}$$

Proof: Let $\widetilde{\mathbf{y}}_\ell = [y_1, \ldots, y_n, \ell]^T$, $\mathbf{a}_u = [K(x_1, u), \ldots, K(x_n, u)]^T$ and $b = K(u, u) = 1$. Let $\mathbf{K}$ be the kernel matrix for the elements in $\mathcal{L}$ and $\widetilde{\mathbf{K}}_u$ be the kernel matrix for the elements in $\mathcal{L} \cup \{u\}$, as defined in (12). Then, for $\ell \in \{-1, +1\}$
$$\begin{aligned} \left\|f_\ell^u(x)\right\|_{\mathcal{H}}^2 &= \widetilde{\mathbf{y}}_\ell^T \, \widetilde{\mathbf{K}}_u^{-1} \, \widetilde{\mathbf{y}}_\ell \overset{(a)}{=} \mathbf{y}^T \left(\mathbf{K} - \mathbf{a}_u \mathbf{a}_u^T\right)^{-1} \\ &\quad \times \mathbf{y} - 2\ell \, \mathbf{y}^T \left(\mathbf{K} - \mathbf{a}_u \mathbf{a}_u^T\right)^{-1} \mathbf{a} + \left(1 - \mathbf{a}_u^T \mathbf{K}^{-1} \mathbf{a}_u\right)^{-1} \\ &\overset{(b)}{=} \mathbf{y}^T \, \mathbf{K}^{-1} \, \mathbf{y} + \frac{\left(1 - \ell \, \mathbf{y}^T \, \mathbf{K}^{-1} \mathbf{a}_u\right)^2}{1 - \mathbf{a}_u^T \mathbf{K}^{-1} \mathbf{a}_u} \overset{(c)}{=} \|f(x)\|_{\mathcal{H}}^2 \\ &\quad + \frac{\left[1 - \ell f(u)\right]^2}{1 - \mathbf{a}_u^T \mathbf{K}^{-1} \mathbf{a}_u}. \end{aligned}$$

where Schur's complement formula gives (a) and Woodbury Identity with some algebra gives (b). We are using the property that $K(x,x)=1$ and the diagonal elements of matrix $\widetilde{\mathbf{K}}_u$ are equal to one. (c) uses (10) for the minimum norm interpolating function based on $\mathcal{L}$, i.e., $f(x)$. Hence, $\|f^u_+(x)\|_{\mathcal{H}} > \|f^u_-(x)\|_{\mathcal{H}}$ if and only if $f(u) < 0$ which gives the statement of proposition. ■

C. Radial Basis Kernels

From here on, we will focus on minimum norm interpolating functions with radial basis kernels. The kernel functions we use have the following form: For $x, x' \in \mathbb{R}^d$, $h > 0$ and $p > 1$, let

$$k_{h,p}(x,x') = \exp\left(-\frac{1}{h}\|x-x'\|_p\right), \tag{13}$$

where $\|x\|_p := (\sum_{i=1}^d x_i^p)^{1/p}$ is the ℓ_p norm and $\|x-x'\|_p$ is the Minkowski distance satisfying the triangle inequality. For $p=1,2$ this category of kernels construct Reproducing Kernel Hilbert Spaces. When the parameters h and p are specified, we denote the kernel function $k_{h,p}(x,y)$ by $k(x,y)$.

D. Laplace Kernel in One Dimension

To develop some intuition, we consider active learning in one-dimension. The sort of target function we have in mind is a multiple threshold classifier. Optimal active learning in this setting coincides with binary search. We now show that the proposed selection criterion based on $\mathsf{score}_{\mathcal{H}}$ with Hilbert norm associated with the Laplace kernels result in an optimal active learning in one dimension (proof in Appendix B.1).

Proposition 2 (Maximin Criteria in One Dimension With Laplace Kernel): Define $K(x,x') = \exp(-|x-x'|/h)$ to be the Laplace kernel in one dimension and the minimum norm interpolator function defined in Section IV-A. Let the selection criterion be based on $\mathsf{score}_{\mathcal{H}}(u)$ function defined in (6) with the Laplace kernel Hilbert norm. Then the following statements hold for any value of $h > 0$:

1) Let x_1 and x_2 be two neighboring labeled points in $\mathcal{L}$. Then $\mathsf{score}_{\mathcal{H}}(\frac{x_1+x_2}{2}) \geq \mathsf{score}_{\mathcal{H}}(u)$ for all $x_1 < u < x_2$.
2) Let $x_1 < x_2$ and $x_3 < x_4$ be two pairs of neighboring labeled points such that $x_2 - x_1 \geq x_4 - x_3$, then
   - if $y_1 \neq y_2$ and $y_3 = y_4$. Then $\mathsf{score}_{\mathcal{H}}(\frac{x_1+x_2}{2}) \geq \mathsf{score}_{\mathcal{H}}(\frac{x_3+x_4}{2})$.
   - if $y_1 = y_2$ and $y_3 \neq y_4$. Then $\mathsf{score}_{\mathcal{H}}(\frac{x_1+x_2}{2}) \leq \mathsf{score}_{\mathcal{H}}(\frac{x_3+x_4}{2})$.
   - if $y_1 \neq y_2$ and $y_3 \neq y_4$. Then $\mathsf{score}_{\mathcal{H}}(\frac{x_1+x_2}{2}) \leq \mathsf{score}_{\mathcal{H}}(\frac{x_3+x_4}{2})$.
   - if $y_1 = y_2$ and $y_3 = y_4$. Then $\mathsf{score}_{\mathcal{H}}(\frac{x_1+x_2}{2}) \geq \mathsf{score}_{\mathcal{H}}(\frac{x_3+x_4}{2})$.

The key conclusion drawn from these properties is that the midpoints between the closest oppositely labeled neighboring examples have the highest score. If there are no oppositely labeled neighbors, then the score is largest at the midpoint of the largest gap between consecutive samples. Thus, the score results in a binary search for the thresholds defining the classifier. Using the proposition above, it is easy to show the following result, proved in the Appendix B.3.

Corollary 2: Consider N points uniformly distributed in the interval $[0,1]$ labeled according to a k-piecewise constant function $g(x)$ so that $y_i = g(x_i) \in \{-1,+1\}$ and length of the pieces are roughly on the order of $\Theta(1/K)$. Then by running the proposed active learning algorithm with Laplace Kernel and any bandwidth, after $O(k \log N)$ queries the sign of the resulting interpolant f correctly labels all N examples (i.e., the training error is zero).

This statement is true for $N > 5/h$. The proof is provided in Appendix B.3.

E. General Radial-Basis Kernels in One Dimension

In the next proposition, we look at the special case of radial basis kernels, defined in Equation(13) applied to one dimensional functions with only three initial points. We show how maximizing $\mathsf{score}_{\mathcal{H}}$ with the appropriate Hilbert norm is equivalent to picking the zero-crossing point of our current interpolator.

Proposition 3 (One Dimensional Functions With Radial Basis Kernels): Assume that for any pair of samples $x, x' \in \mathcal{L}$ we have $|x-x'| \geq \Delta$. Assume $\Delta h^{-1/p} \geq D$ for a constant value of D. Let $x_1 < x_2 < x_3 \in \mathbb{R}$, $y_1 = y_2 = +1$ and $y_3 = -1$. For u such that $x_2 + \Delta/2 < u < x_3 - \Delta/2$, we have $\mathsf{score}_{\mathcal{H}}(u) \leq \mathsf{score}_{\mathcal{H}}(u^*)$ where u^* is the point satisfying $f(u^*) = 0$.

The proof is rather tedious and appears in Appendix C.1. But the idea is based on showing that with small enough bandwidth, $\|f^u_+\|$ is increasing in u in the interval $[x_2 + \Delta/2, x_3 - \Delta/2]$ and $\|f^u_-\|$ is decreasing in u in the same interval. This shows that $\max_u \min_{\ell \in \{-1,+1\}} \|f^u_\ell\|$ occurs at u^* such that $\|f^{u^*}_+\| = \|f^{u^*}_-\|$. We showed that this is equivalent to the condition $f(u^*) = 0$.

F. Properties of Data Based-Norm Criterion

Intuitively, $\mathsf{score}_{\mathrm{D}}$ measures the expected change in the squared norm over all unlabeled examples if $u \in \mathcal{U}$ is selected as the next point. This norm is sensitive to the particular distribution of the data, which is important if the data are clustered. This behavior will be demonstrated in the multidimensional setting discussed next.

In this section, we present two theoretical results on the properties of data-based norm selection criterion. To do so, we will prove the properties of the selected examples based on the data-based norm in the context of the clustered data. In particular, if the support of the generative distribution $P_X(x)$ is composed of several disjoint clusters, the data-based norm criterion prioritizes labeling samples from bigger clusters first. Subsequently, it selects a sample from each cluster to be labeled. If the clustering in the dataset is aligned with their labels (most of the samples in the same cluster are in the same class), labeling one sample in each cluster ensures rapid decay in the probability of error of the classifier as a function of number of labeled samples. This behavior is consistent with numerical simulations presented in Section V.

The next theorem will show that if the clusters are well-separated (the distance between the clusters are sufficiently

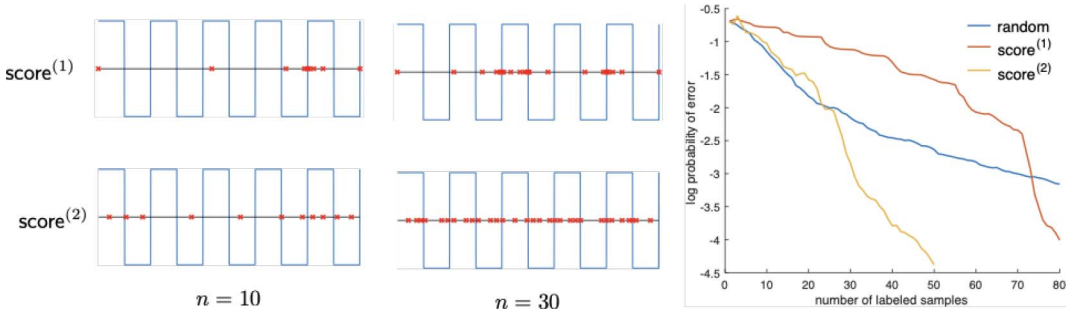

Fig. 2. Uniform distribution of samples in unit interval, multiple thresholds between ±1 labels, and active learning using Laplace Kernel, Bandwidth= 0.1. Probability of error of the interpolated function shown on right.

large), then the first example to be selected to for labeling is in the biggest cluster.

Theorem 4 (First Point in Clustered Data): Fix $p > 1$ and $h > 0$. Let the distribution $\mathbb{P}(X)$ be uniform over M disjoint sets $B_1, \ldots, B_M$ such that B_i is an ℓ_p ball with radius r_i and center c_i, i.e.,

$$B_i = \mathcal{B}_{d,p}(r_i; c_i) := \Big\{x \in \mathbb{R}^d : \|x - c_i\|_p \le r_i\Big\}. \qquad (14)$$

Without loss of generality, assume $r_1 > r_2 > \cdots > r_K$. Define $D = \min_{i\neq j} \|c_i - c_j\|_p - 2r_1$ as an upper bound for the minimum distance between the clusters.

Assume $\mathcal{L} = \varnothing$ and let the interpolating functions f be defined in (10) with $k_{h,p}$ (defined in (13)). The selection criterion is based on the $\mathsf{score}_{\mathrm{D}}$ function defined in (7). If

$$D > \frac{h}{2}\Big[\ln M - \ln\Big(1 - (r_2/r_1)^d\Big)\Big] \qquad \text{and} \qquad r_1 \le h/2,$$

then the first point to be labeled is in the biggest ball, B_1.

The proof is presented in Appendix C.1.

The next theorem shows that if the distance between the clusters are sufficiently large and the radius of the clusters are not too large, then the active learning algorithm based on the notion of score with data-based norm labels one sample from each cluster before zooming in inside the clusters.

Theorem 5 (Cluster Exploration): Let $\mathcal{S}$ be the support of P_X. Assume $\mathcal{S} = \cup_{i=1}^{M} B_i$ where B_i's are ℓ_p-balls with radii r and centers c_i. Define $D := \min_{i\neq j} \|c_i - c_j\|_p - 2r_1$ to be the minimum distance between the clusters. Let $\mathcal{L} = \{x_1, x_2, \ldots, x_L\}$ be $L < M$ labeled points such that $x_1 \in B_1, x_2 \in B_2, \ldots, x_L \in B_L$. Let the selection criterion be based on the $\mathsf{score}_{\mathrm{D}}$ function defined in (7). If $r < h/3$ and $D \ge 12h\ln(2M)$, then the next point to be labeled is in a new ball $(\cup_{i=L+1}^{M} B_i)$ containing no labeled points.

As a corollary of the above theorem, one can see that if the ratio of the distance between the clusters to the radius of clusters is sufficiently large $(D/r > 36ln(2M))$, then one can use a kernel with proper bandwidth which picks one sample from each cluster initially. The proof is presented in Appendix C.2.

V. Numerical Simulations of Kernel Based

In this Section, we present the outcome of numerical simulations of the proposed selection criteria on synthetic and real data. In this section, $\mathsf{score}_{\mathcal{H}}$ is used to denoted the score function defined in (6) with the Hilbert norm associated with the Laplace Kernel. Similarly, $\mathsf{score}_{\mathrm{D}}$ is the score function defined in (7) with the data-based norm.

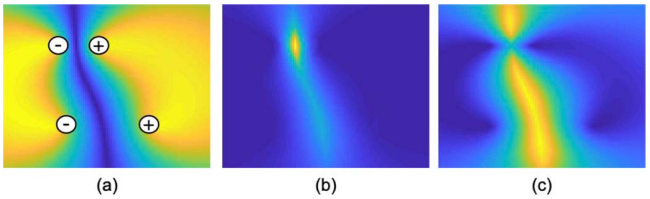

Fig. 3. Data selection of Laplace kernel active learner. (a) Magnitude of output map kernel machine trained to interpolate four data points as indicated (dark blue is 0 indicating the learned decision boundary). (b) Max-Min RKHS norm selection of next point to label. Brightest yellow is location of highest score and selected example. (c) Max-Min selection of next point to label using data-based norm. Both select the point on the decision boundary, but the RKHS norm favors points that are closest to oppositely labeled examples.

A. Bisection in One Dimension

The bisection process is illustrated experimentally in the Figure 2 below. $\mathsf{score}_{\mathcal{H}}$ uses the RKHS norm. For comparison, we also show the behavior of the algorithm using $\mathsf{score}_{\mathrm{D}}$ and the data-based norm. Data selection using either score drives the error to zero faster than random sampling (as shown on the left). We clearly see the bisection behavior of $\mathsf{score}_{\mathcal{H}}$, locating one decision boundary/threshold and then another, as the proof corollary above suggests. Also, we see that the data-based norm does more exploration away from the decision boundaries. As a result, the data-based norm has a faster and more graceful error decay, as shown on the right of the figure. Similar behavior is observed in the multidimensional setting shown in Figure 5.

B. Multidimensional Setting With Smooth Boundary

The properties and behavior found in the one dimensional setting carry over to higher dimensions. In particular, the max-min norm criterion tends to select unlabeled examples near the decision boundary *and* close to oppositely labeled examples, This is illustrated in Figure 3 below. The inputs points (training examples) are uniformly distributed in the square $[-1,\ 1] \times [-1,\ 1]$. We trained an Laplace kernel machine to perfectly interpolate four training points with locations and binary labels as depicted in Figure 3(a). The color depicts the magnitude of the learned interpolating function: dark blue is 0 indicating the "decision boundary" and bright yellow is approximately 3.5.

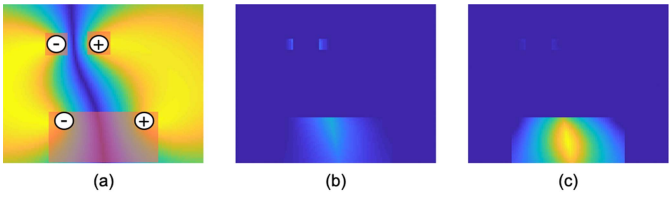

Fig. 4. Data selection of Laplace kernel active learner. (a) Unlabeled examples are only available in magenta shaded regions. (b) Max-Min selection map using RKHS norm (6). (c) Max-Min selection map using data-based norm defined in Equation (7).

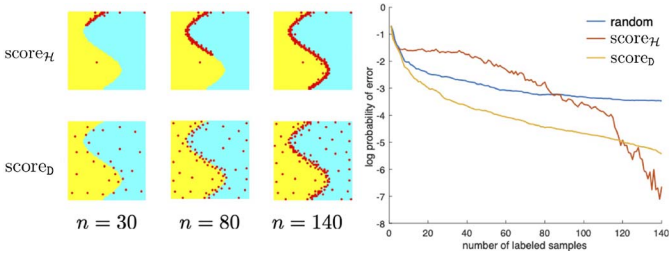

Fig. 5. Uniform distribution of samples, smooth boundary, Laplace Kernel, Bandwidth $= 0.1$. On left, sampling behavior of score$_{\mathcal{H}}$ and score$_{\mathrm{D}}$ at progressive stages (left to right). On right, error probabilities as a function of number of labeled examples.

Figure 3(b) denotes the score for selecting a point at each location based on RKHS norm criterion. Figure 3(c) denotes the score for selecting a point at each location based on data-based norm criterion discussed above. Both criteria select the point on the decision boundary, but the RKHS norm favors points that are closest to oppositely labeled examples whereas the data-based norm favors points on the boundary further from labeled examples.

Next we present a modified scenario in which the examples are not uniformly distributed over the input space, but instead concentrated only in certain regions indicated by the magenta highlights in Figure 4(a). In this setting, the example selection criteria differ more significantly for the two norms. The weight norm selection criterion remains unchanged, but is applied only to regions where there are examples. Areas with out examples to select are indicated by dark blue in Figure 4(b)-(c). The data-based norm is sensitive to the non-uniform input distribution, and it scores examples near the lower portion of the decision boundary highest.

The distinction between the max-min selection criterion using the RKHS vs. data-based norm is also apparent in the experiment in which a curved decision boundary in two dimensions is actively learned using a Laplace kernel machine, as depicted in Figure 5 below. score$_{\mathcal{H}}$ is the max-min RKHS norm criterion at progressive stages of the learning process (from left to right). The data-based norm is used in score$_{\mathrm{D}}$ defined in Equation (7). Both dramatically outperform a passive (random sampling) scheme and both demonstrate how active learning automatically focuses sampling near the decision boundary between the oppositely labeled data (yellow vs. blue). However, the data-based norm does more exploration away from the decision boundary. As a result, the data-based norm requires slightly more labels to perfectly predict all unlabeled examples, but has a more graceful error decay, as shown on the right of the figure.

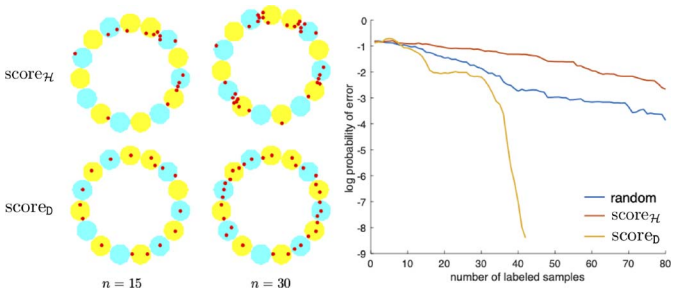

Fig. 6. Uniform distribution of samples, smooth boundary, Laplace Kernel, Bandwidth $= 0.1$. On left, sampling behavior of score$_{\mathcal{H}}$ and score$_{\mathrm{D}}$ at progressive stages (left to right). On right, error probabilities as a function of number of labeled examples.

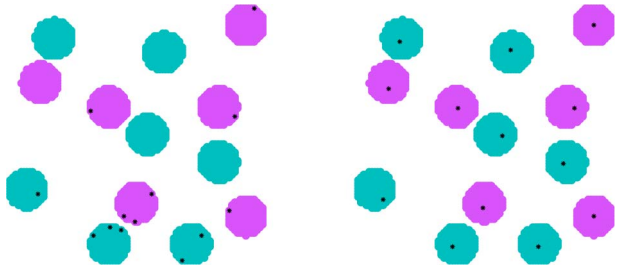

Fig. 7. Points in blue and yellow clusters are labeled $+1$ and -1, respectively. The left figure uses score$_{\mathcal{H}}$ to be the score function defined in (6) with the Hilbert norm associated with the Laplace Kernel. Similarly, score$_{\mathrm{D}}$ is the score function defined in (7) with the data-based norm. The first 13 samples selected by score$_{\mathcal{H}}$ and score$_{\mathrm{D}}$ are depicted as black dots. score$_{\mathrm{D}}$ has labeled one sample from each cluster, but score$_{\mathcal{H}}$ has not labeled any samples from 5 clusters. Note that score$_{\mathcal{H}}$ has spent some of the sample budget to discriminate between nearby clusters with opposite labels.

C. Multidimensional Setting With Clustered Data

To capture the properties of the proposed selection criteria in clustered data, we implemented the algorithm on synthetic clustered data in Figures 6 and 7. We demonstrate how the data-based norm also tends to automatically select representative examples from clusters when such structure exists in the unlabeled dataset. Figure 6 compares the behavior of selection based on score$_{\mathcal{H}}$*In* with the RKHS norm and score$_{\mathrm{D}}$ with data-based norm, when data are clusters and each cluster is homogeneously labeled. We see that the data-based norm quickly identifies the clusters and labels a representative from each, leading to faster error decay as shown on the right.

In the setup in Figure 7, the samples are generated based on a uniform distribution on 13 clusters. Points in blue and yellow clusters are labeled $+1$ and -1, respectively. We run the two variations of proposed active learning algorithms and compare their sampling strategy in this setup. The left figure uses score$_{\mathcal{H}}$ to be the score function defined in (6) with the Hilbert norm associated with the Laplace Kernel. Similarly, score$_{\mathrm{D}}$ is the score function defined in (7) with the data-based norm.

The selection criterion based on score$_{\mathcal{H}}$ prioritizes sampling on the decision boundary of the current classifier where the currently oppositely labeled samples are close to each other. This behavior of the algorithm based on score$_{\mathcal{H}}$ in one dimension is proved in Sections IV-D and IV-E. Alternatively, score$_{\mathrm{D}}$ prioritizes labeling at least one sample from each

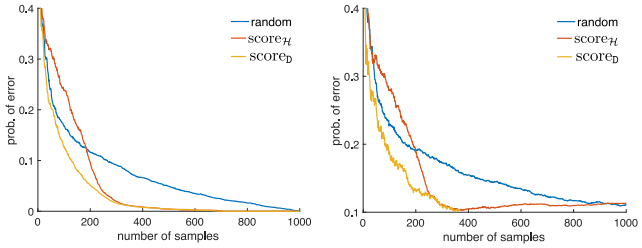

Fig. 8. Probability of error for learning a classification task on MNIST data set. The performance of three selection criteria for labeling the samples: random selection, active selection based on $\mathsf{score}_{\mathcal{H}}$, and active selection based on $\mathsf{score}_{\mathrm{D}}$. The first curve depicts the probability of error on the training set and the second curve is the probability of error on the test set.

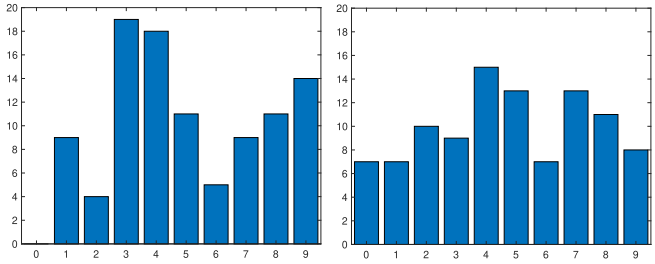

Fig. 9. The histogram of the handwritten digits associated with the labeled samples after labeling 100 samples. The first histogram is for the selection criterion $\mathsf{score}_{\mathcal{H}}$ and the second histogram is for the selection criterion $\mathsf{score}_{\mathrm{D}}$. Notably, $\mathsf{score}_{\mathcal{H}}$ has not labeled any of the images of the digit 0.

cluster. Hence, after labeling 13 samples, the active learning algorithm based on $\mathsf{score}_{\mathrm{D}}$ has one sample in each cluster, but the active learning algorithm based on $\mathsf{score}_{\mathrm{D}}$ has not labeled any samples in 5 clusters.

D. MNIST Experiments

Here we illustrate the performance of the proposed active learning method on the MNIST dataset. We ran algorithms based on our proposed selection criteria for a binary classification task on MNIST dataset. The binary classification task used in this experiment assigns a label -1 to any digit in set $\{0, 1, 2, 3, 4\}$ and label $+1$ to $\{5, 6, 7, 8, 9\}$. The goal of the classifier is detecting whether an image belongs to the set of numbers greater or equal to 5 or not. We used Laplace kernel as defined in (13) with $p = 2$ and $h = 10$ on the vectorized version of a dataset of 1000 images. In Figures 8, $\mathsf{score}_{\mathcal{H}}$ is the score function defined in (6) with the Hilbert norm associated with the Laplace Kernel. Similarly, $\mathsf{score}_{\mathrm{D}}$ is the score function defined in (7) with the data-based norm.

To asses the quality of performance of each of the selection criteria, we compare the probability of error of the interpolator at each iteration. In particular, we plot the probability of error of the interpolator as a function of number of labeled samples, using the $\mathsf{score}_{\mathcal{H}}$ and $\mathsf{score}_{\mathrm{D}}$ functions on the training set and test set separately. For comparison, we also plot the probability of error when the selection criterion for picking samples to be labeled is random.

Figure 8 (a) shows the decay of probability of error in the training set. When the number of labeled samples is equal to the number of samples in the training set, it means that all the samples in training set are labeled and used in constructing the interpolator. Hence, the probability of error on the training set for any selection criterion is zero when number of labeled samples is equal to the number of samples in the training set. Figure 8 (b) shows the probability of error on the test set as a function of the number of labeled samples in the training set selected by each selection criterion.

1) Clustering in MNIST: The binary classification task used in the MNIST experiment assigns a label -1 to any digit in set $\{0, 1, 2, 3, 4\}$ and label $+1$ to $\{5, 6, 7, 8, 9\}$. We expect that the images are clustered where each cluster would correspond to the images of a digit. We expect that the advantageous behavior of using data-based norm criterion in clustered data is one of the reasons for faster decay of probability of error of the $\mathsf{score}_{\mathrm{D}}$ in Figure 8.

To verify this intuition, we look at the samples that were chosen by each criterion and the digit corresponding to that sample. Note that this digit is the number represented in the image and not the label of the sample since the label of each sample is $+1$ or -1 depending whether the number is greater than 4 or not. After labeling 100 samples, we look at histogram of the digits associated with the labeled samples with each criterion $\mathsf{score}_{\mathcal{H}}$ and $\mathsf{score}_{\mathrm{D}}$. If samples of each cluster are chosen to be labeled uniformly among clusters, we would see about 10 labeled samples in each cluster. Figure 9 shows the histogram described above for two variations of the selection criteria based on $\mathsf{score}_{\mathcal{H}}$ or $\mathsf{score}_{\mathrm{D}}$. We observe that selecting samples based on $\mathsf{score}_{\mathrm{D}}$ is much more uniform among the clusters. On the contrary, selecting samples based on $\mathsf{score}_{\mathcal{H}}$ gives much less uniform samples among clusters. In the particular example given in Figure 9, we see that even after selecting 100 samples to be labeled, no sample in the cluster of images of number 0 has been labeled in this instance of execution of the selection algorithm based on notion of $\mathsf{score}_{\mathcal{H}}$.

To quantify the uniformity of selecting samples in different clusters, we ran this experiment 20 times and estimated the standard deviation of number of labeled samples in each cluster after labeling 100 samples. Note that since we have 10 clusters, the mean of the number of labeled samples in each cluster is 10. The standard deviation using $\mathsf{score}_{\mathcal{H}}$ is 4.1 whereas standard deviation using $\mathsf{score}_{\mathcal{H}}$ is 2.7. This shows that selection criterion based on $\mathsf{score}_{\mathrm{D}}$ samples more uniformly among the clusters.

VI. Interpolating Neural Network Active Learners

Here we briefly examine the extension of the max-min criterion and its variants to neural network learners. Neural network complexity or capacity can be controlled by limiting magnitude of the network weights [24], [25], [26]. A number of weight norms and related measures have been recently proposed in the literature [27], [28], [29], [30], [31]. For example, ReLU networks with a single hidden layer and minimum ℓ_2 norm weights coincide with linear spline interpolation [32]. With this in mind, we provide empirical evidence showing that defining the max-min criterion with the norm of the network weights yields a neural network active learning algorithm with properties analagous to those obtained in the RKHS setting.

Consider a single hidden layer network with ReLU activation units trained using MSE loss. In Figure 10 we show the

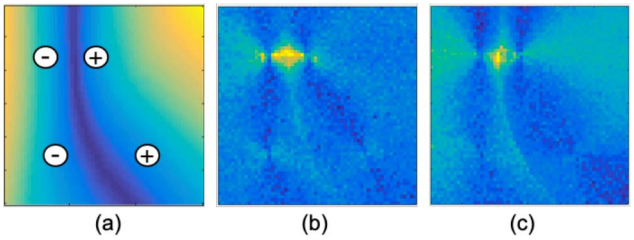

Fig. 10. Data selection of neural network active learner. (a) Magnitude of output map of single hidden layer ReLU network trained to interpolate four data points as indicated (dark blue is 0 indicating the learned decision boundary). (b) Max-Min selection of next point to label using network weight norm. (c) Max-Min selection of next point to label using data-based norm. Both select the point on the decision boundary that is closest to oppositely labeled examples.

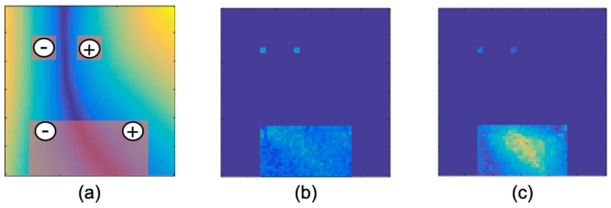

Fig. 11. Data selection of neural network active learner. (a) Unlabeled examples are only available in magenta shaded regions. (b) Max-Min selection map using network weight norm. (c) Max-Min selection map using data-based norm.

results of an experiment implemented in PyTorch in the same settings considered above for kernel machines in Figures 3 and 4. We trained an overparameterized network with 100 hidden layer units to perfectly interpolate four training points with locations and binary labels as depicted in Figure 10(a). The color depicts the magnitude of the learned interpolating function: dark blue is 0 indicating the "decision boundary" and bright yellow is approximately 3.5. Figure 10(b) denotes the $\mathsf{score}_{\mathcal{H}}$ with the weight norm (i.e., the ℓ_2 norm of the resulting network weights when a new sample is selected at that location). The brightest yellow indicates the highest score and the location of the next selection. Figure 10(c) denotes the $\mathsf{score}_{\mathrm{D}}$ with the data-based norm defined in Equation (7). In both cases, the max occurs at roughly the same location, which is near the current decision boundary *and* closest to oppositely labeled points. The data-based norm also places higher scores on points further away from the labeled examples. Thus, the data selection behavior of the neural network is analagous to that of the kernel-based active learner (compare with Figure 3).

Next we present a modified scenario in which the examples are not uniformly distributed over the input space, but instead concentrated only in certain regions indicated by the magenta highlights in Figure 11(a). In this setting, the example selection criteria differ more significantly for the two norms. The weight norm selection criterion remains unchanged, but is applied only to regions where there are examples. Areas without examples to select are indicated by dark blue in Figure 11(b)-(c). The data-based norm is sensitive to the non-uniform input distribution, and it scores examples near the lower portion of the decision boundary highest. Again, this is quite similar to the behavior of the kernel active learner (compare with Figure 4).

VII. Conclusion and Future Work

The question of designing active learning algorithms in the regime of nonparametric and overparameterized models become more essential as we look at larger models which require bigger training sets. To reduce the human cost of labeling allyl samples, we can use a pool-based active learning algorithm to avoid labeling non-informative examples.

Our algorithm does not exploit any assumption about the underlying classifier in selecting the samples to label. Yet, for a wide range of classifiers, it performs well with provable guarantees. It is designed for the extreme case of the nonparametric setting in which no assumption about the smoothness of the boundary between different classes is made by the learner.

There are many interesting questions remaining: the behavior of our proposed criterion applied to other classifiers such as kernel SVM instead of minimum norm interpolators, generalization of the criterion to multi-class settings and regression algorithms. The computational complexity of our criterion can also be a serious bottleneck in applications with bigger data-sets and should be addressed in future. Additional numerical simulations, especially with more complex architecture of Neural Networks can also be insightful.

References

[1] S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 3331–3340.

[2] M. Belkin, S. Ma, and S. Mandal, "To understand deep learning we need to understand kernel learning," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 540–548.

[3] S. Vaswani, F. Bach, and M. Schmidt, "Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron," 2018. [Online]. Available: arXiv:1810.07288.

[4] M. Belkin, A. Rakhlin, and A. B. Tsybakov, "Does data interpolation contradict statistical optimality?" 2018. [Online]. Available: arXiv:1806.09471.

[5] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine learning and the bias-variance trade-off," 2018. [Online]. Available: arXiv:1812.11118.

[6] S. Arora, S. S. Du, W. Hu, Z. Li, and R. Wang, "Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks," 2019. [Online]. Available: arXiv:1901.08584.

[7] M. Belkin, D. J. Hsu, and P. Mitra, "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 2300–2311.

[8] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," 2019. [Online]. Available: arXiv:1903.08560.

[9] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," 2019. [Online]. Available: arXiv:1903.07571.

[10] T. Liang and A. Rakhlin, "Just interpolate: Kernel 'ridgeless' regression can generalize," 2018. [Online]. Available: arXiv:1808.00387.

[11] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *J. Mach. Learn. Res.*, vol. 2, pp. 45–66, Nov. 2001.

[12] D. A. Cohn, L. E. Atlas, and R. E. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.

[13] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," 2017. [Online]. Available: arXiv:1708.00489.

[14] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian active learning with image data," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1183–1192.

[15] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2591–2600, Jan. 2017.

[16] Y. Shen, H. Yun, Z. C. Lipton, Y. Kronrod, and A. Anandkumar, "Deep active learning for named entity recognition," 2017. [Online]. Available: arXiv:1707.05928.
[17] B. Settles, "Active learning literature survey," Dept. Comput. Sci., Univ. Wisconsin-Madison, Madison, WI, USA, Rep. 1648, 2009.
[18] B. Settles, "Active learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 6, no. 1, pp. 1–114, 2012.
[19] S. Dasgupta, "Two faces of active learning," *Theor. Comput. Sci.*, vol. 412, no. 19, pp. 1767–1781, 2011.
[20] P. H. P. Savarese, I. Evron, D. Soudry, and N. Srebro, "How do infinite width bounded norm networks look in function space?" in *Proc. Conf. Learn. Theory (COLT)*, Phoenix, AZ, USA, 2019, pp. 2667–2690.
[21] R. Parhi and R. D. Nowak, "Minimum 'norm' neural networks are splines," 2019. [Online]. Available: arXiv:1910.02333.
[22] P. L. Bartlett, "The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network," *IEEE Trans. Inf. Theory*, vol. 44, no. 2, pp. 525–536, Mar. 1998.
[23] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta numerica*, vol. 8, pp. 143–195, Jan. 1999.
[24] P. L. Bartlett, "For valid generalization the size of the weights is more important than the size of the network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1997, pp. 134–140.
[25] B. Neyshabur, R. Tomioka, and N. Srebro, "In search of the real inductive bias: On the role of implicit regularization in deep learning," 2014. [Online]. Available: arXiv:1412.6614.
[26] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," 2016. [Online]. Available: arXiv:1611.03530.
[27] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Proc. Conf. Learn. Theory*, 2015, pp. 1376–1401.
[28] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, "Spectrally-normalized margin bounds for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6240–6249.
[29] N. Golowich, A. Rakhlin, and O. Shamir, "Size-independent sample complexity of neural networks," in *Proc. Conf. Learn. Theory*, 2018, pp. 297–299.
[30] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 254–263.
[31] B. Neyshabur, S. Bhojanapalli, and N. Srebro, "A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks," 2017. [Online]. Available: arXiv:1707.09564.
[32] P. Savarese, I. Evron, D. Soudry, and N. Srebro, "How do infinite width bounded norm networks look in function space?" 2019. [Online]. Available: arXiv:1902.05040.