2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)
Allerton Park and Retreat Center
Monticello, IL, USA, September 24-27, 2019

Maximin Active Learning with Data-Dependent Norms

Mina Karzand
Wisconsin Institute of Discovery
University of Wisconsin, Madison
Madison, WI
karzand@wisc.edu

Robert D. Nowak
Electrical Engineering Department
University of Wisconsin, Madison
Madison, WI
rdnowak@wisc.edu

***Abstract*—Overparameterized machine learning models are often fit perfectly to training data, yet remarkably generalize well to new data. However, learning good models can require an enormous number of labeled training data. This challenge motivates the study of active learning algorithms that sequentially and adaptively request labels for "informative" examples for a large pool of unlabeled data. A maximin criterion was recently proposed for active learning specifically in the overparameterized and interpolating regime. Roughly speaking, the maximin criterion selects the example that is most difficult to interpolate, as measured by an appropriate norm on the interpolating function. Data-dependent norms perform best empirically, exhibiting intriguing adaptivity to cluster structure within the data. The main contribution of this paper is to mathematically characterize this behavior. Our main results show that the maximin criterion based on data-dependent norms provably discovers clusters and also automatically generates labeled coverings of the dataset.**

***Index Terms*—Active Learning, Reproducing Hilbert Kernel Spaces, Data-dependent Norm**

I. Introduction

Deep neural networks have revolutionized machine learning applications, and theoreticians have struggled to explain their surpising properties. Deep neural networks 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], [4], [5], [7], [10], [13], [19], which 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 even shallow (but wide) networks and classical kernel methods fit to the data but regularized to small weights (e.g., minimum norm fit to data) can generalize well [3], [5], [10], [12].

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 class labels, for example, can be prohibitive in many applications. This has prompted researchers to investigate active learning for neural networks [8], [9], [11], [14], [17], [20]. Active learning algorithms have access to a large but unlabeled dataset of examples and sequentially select the most "informative" examples for labeling [15], [16] . This can reduce the total number of labeled examples needed to learn an accurate model.

This work was partially supported by the Air Force Machine Learning Center of Excellence FA9550-18-1-0166.

This paper builds on a new framework for active learning in the overparameterized and interpolationg regime [11], focusing on kernel methods (which can be viewed as single hidden-layer networks). That work proposed an active learning algorithm based on the notion of minimum norm interpolators. The algorithm selects examples to label based on a maximin criterion. Roughly speaking, the maximin criterion selects 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.

In [11], it is shown that the maximin criterion using the RKHS norm tends to select unlabeled examples near the decision boundary and close to oppositely labeled examples, allowing the algorithm to focus on learning decision boundaries. The maximin criterion reduces to optimal binary search in the one-dimensional linear classifier setting, and several other interesting properties were established for the criterion. Experimentally, it was shown that using a data-based norm in the "max" step (instead of the RKHS norm) exhibits the desirable behavior of automatically discovering cluster structure in unlabeled data and labeling representative examples from each cluster. The main contribution of this paper is to mathematically characterize the clustering behavior. Our main results show that the maximin criterion provably discovers clusters and also automatically generates labeled coverings of the dataset.

II. Selection Criterion

In [11], we introduced a selection process for active learning algorithms. According to this process, at each time step, the algorithm has access to a pool of labeled samples $\mathcal{L} = \{(x_1, y_1), \cdots, (x_L, y_L)\}$ and a set of unlabaled samples $\mathcal{U}$ where $x_i \in \mathcal{X}$, $y_i \in \{-1, +1\}$ for all $i \leq L$ and $\mathcal{U} \subseteq \mathcal{X}$. 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.

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

Note that for any unlabeled sample $u \in \mathcal{U}$, the value of $\mathsf{score}(u)$ depends implicitly on the set of currently labeled points, $\mathcal{L}$: information gained by labeling u depends on the current knowledge of the learner, *i.e.*, $\mathcal{L}$. To define our proposed notion of score, we introduce some notations next.

Given the set of labeled samples, $\mathcal{L}$, and a set of functions $\mathcal{F}$ mapping $\mathcal{X}$ to $\mathbb{R}$, let $f(x) \in \mathcal{F}$ be the interpolating function such that $f(x_i) = y_i$ for all $(x_i, y_i) \in \mathcal{L}$. Clearly, the definition of $f(x)$ depends on the set of currently labeled samples $\mathcal{L}$, although we omit this dependency from the notation for the sake of brevity. Also, note that there are many functions that interpolate a discrete set of points such as $\mathcal{L}$. We define $f(x)$ to be the minimum norm interpolator.

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

This definition requires definition of norm associated with the function class $\mathcal{F}$. The choice of $\mathcal{F}$ and the norm $\|\cdot\|_{\mathcal{F}}$ is application dependent. In this paper, we look into function classes in reproducing kernel Hilbert spaces where the norm is the corresponding Hilbert norm.

Roughly speaking, for a sample $u \in \mathcal{U}$, we want $\mathsf{score}(u)$ to measure the amount of change in the interpolating function upon labeling u. For $t \in \{-1, +1\}$ and $u \in \mathcal{U}$, define $f_u^t(x)$ to be a new minimum norm interpolating function based on current set of labeled samples $\mathcal{L}$, their labels and sample u with label t:

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

So, $\mathsf{score}(u)$ should reflect $\|f_u^t(x) - f(x)\|$ which is exactly the average magnitude of the change in the interpolating function upon labeling u with t.

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 $t(u) \in \{-1, +1\}$ and compute $\mathsf{score}(u)$ assuming that upon labeling, u will be labeled $t(u)$. Similar to [11], we propose the following criterion for choosing $t(u)$:

$$t(u) := \text{argmin}_{t \in \{-1,+1\}} \|f_u^t(x)\|_{\mathcal{F}}\,. \tag{4}$$

We are estimating the label of any unlabeled sample, u, to be the one which gives the smoothest interpolating function among the two possible functions $f_u^+(x)$ and $f_u^-(x)$.

In this paper, we will focus on interpolating functions in a Reproducing Kernel Hilbert space (RKHS). In [11], it is proved that when $f(x)$ is constrained to be an element of RKHS and the norm used in (4) is the Hilbert norm, we have

$$t(u) = \begin{cases} +1 & \text{if } f(u) \geq 0 \\ -1 & \text{if } f(u) < 0 \end{cases}\,.$$

We use this property about the estimated label of sample u denoted by $t(u)$ in this paper.

Define

$$f_u(x) := f_u^{t(u)}(x)$$

to be the interpolating function after adding the sample u with the label $t(u)$, defined above. We define the notion of score as

$$\mathsf{score}(u) := \|f_u(x) - f(x)\|\,. \tag{5}$$

The theoretical results in [11] focuses on interpolating functions in RKHS and the score function based on the Hilbert norm associated with RKHS. Several theoretical guarantees provide the intuition that the score function based on the Hilbert norm focuses on labeling the samples that are near the decision boundary of the current interpolator, and are close to the oppositely labeled samples. This intuition is confirmed in the numerical simulations presented in that paper.

An alternative approach presented in [11] is using norms that are different from Hilbert norm in the definition of the score function (5). The *data-based norm* is defined as

$$\|f_u(x) - f(x)\|_{\mathcal{U}} := \sqrt{\frac{1}{|\mathcal{U}|} \sum_{x \in \mathcal{U}} \big[f_u(x) - f(x)\big]^2}\,. \tag{6}$$

This notion of score function measures the average magnitude of change in the interpolating function, evaluated at the unlabeled samples. The intuition behind this definition is the following: If the samples are generated independently and identically based on the distribution $P_X(x)$, and if the pool of unlabeled samples is dense, then the above sum is an estimation of the following term

$$\sqrt{\mathbb{E}_{P_X} \big[f_u(x) - f(x)\big]^2}\,. \tag{7}$$

So score function would give larger weight to changes in the interpolating functions in the areas of input space $\mathcal{X}$ which have higher probability. Similar data-based norms have been proposed in the standard, passive learning setting [6], [18].

One particular example of this sensitivity to the generative distribution P_X is when the data is clustered. This implies that the distribution P_X is nonzero only in union of several compact closed subsets of $\mathcal{X}$. Now, if labeling a certain point changes the interpolating function drastically in the region in which P_X is zero, and does not make much difference in the region in the support of P_X, then labeling this point would not improve the performance of the subsequent classification task significantly: the only change in the decision boundary occurs in the region in which there are no samples in this scenario.

Numerical simulations in [11] show that the active learning algorithm based on a selection criterrio with the notion of score with the data-based norm achieves more graceful decay

of the probability of error compared to the notion of score based on the Hilbert norm. However, there are no theoretical guarantees on the performance of this criterion. In this paper, we provide several theoretical statements on the performance of the selection criterion based on the data-based norm and support these results with various numerical simulations. For our theoretical analysis, we assume that P_X is uniform over its support, denoted by $\mathcal{X}$. Since the selection process of the next sample to be labeled, defined in (1) only depends on the relative value of score function for various unlabeled points, applying a monotonic function on score does not change the outcome of selection process. Hence, we use the following definition of the score function in our theoretical analysis instead of (7).

$$\mathsf{score}(u) := \int_{x\in\mathcal{X}} \big[f_u(x) - f(x)\big]^2 \,\mathrm{d}P_X(x)\,. \tag{8}$$

This is roughly equivalent to the selection criterion based on the empirical data-based norm in (6) in the large-sample limit.

III. Reproducing Kernel Hilbert Space

We will focus on kernel based interpolating functions. The kernel functions we use have the following form: For $x, y \in \mathbb{R}^d$ and $p > 1$

$$k_{h,p}(x,y) = \exp\Big(-\frac{1}{h}\|x-y\|_p\Big)\,, \tag{9}$$

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

For the set of labeled samples $\mathcal{L} = \{(x_1,y_1),\cdots,(x_L,y_L)\}$ with labels in $\{-1,+1\}$, let the function $f(x)$ be decomposed as

$$f(x) = \sum_{i=1}^{L} \alpha_i k(x_i,x) \qquad \text{with} \quad \alpha = \mathbf{K}^{-1}\mathbf{y}, \tag{10}$$

where $\mathbf{K} = \big[\mathbf{K}_{i,j}\big]_{i,j}$ is the L by L matrix such that $\mathbf{K}_{i,j} = k(x_i,x_j)$ and $\mathbf{y} = [y_1,\cdots,y_L]^T$. Using reproducible kernels imply $f(x) \in \mathcal{H}$ for the a RKHS $\mathcal{H}$. Then, $f(x)$ defined above is the minimum Hilbert norm interpolating function defined in (2).

For $u \in \mathcal{U}$ and $t \in \{-1,+1\}$, the minimum norm interpolating unction $f_u^t(x)$ (based on currently labeled samples $\mathcal{L}$ and sample u with label t) is defined similarly :

$$f_u^t(x) = \sum_{i=1}^{L} \widetilde{\alpha}_i k(x_i,x) + \widetilde{\alpha}_{L+1} k(u,x) \qquad \text{with} \quad \widetilde{\alpha} = \widetilde{\mathbf{K}}_u^{-1}\widetilde{\mathbf{y}}_u, \tag{11}$$

where

$$\widetilde{\mathbf{K}}_u = \begin{bmatrix} \mathbf{K} & \mathbf{a}_u \\ \mathbf{a}_u^T & 1 \end{bmatrix}\,,\ \mathbf{a}_u = \begin{bmatrix} k(x_1,u) \\ \vdots \\ k(x_L,u) \end{bmatrix} \text{ and } \widetilde{\mathbf{y}}_t = \begin{bmatrix} \mathbf{y} \\ t \end{bmatrix}\,. \tag{12}$$

IV. Properties of data based-norm criterion

In this section, we present two theoretical results on the properties of data-based norm selection criterion. 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 distance between the clusters are sufficiently large, then the first example to be selected to be labeled is in the biggest cluster.

Theorem 1 (First point in clustered data)**.** *Fix $p > 1$ and $h > 0$. Let the distribution $P(X)$ be uniform over M disjoint sets $B_1,\cdots,B_M$ such that B_i is an ℓ_p ball with radius r_i and center c_i, i.e.,*

$$B_i = \mathcal{B}_{d,p}(r_i;c_i) := \{x \in \mathbb{R}^d : \|x-c_i\|_p \le r_i\}\,. \tag{13}$$

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

Assume $\mathcal{L} = \varnothing$ and we usethe interpolating functions f defined in (10) *with $k_{h,p}$ (defined in* (9)*). The selection criterion is based on the* score *function defined in* (8)*. If*

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

and $r_1 \le h/2$, then the first point to be labeled is in the biggest ball, B_1.

The next theorem will show that if the distance between the clusters are sufficiently large and the radius of the clusters are sufficiently small, 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 2 (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 $\mathcal{L} = \{x_1,x_2,\cdots,x_L\}$ be $L < M$ labeled points such that $x_1 \in B_1, x_2 \in B_2,\cdots,x_L \in B_L$.*

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 ration 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.

V. Numerical Simulations

In this Section, we present the outcome of numerical simulations of the proposed selection criteria on synthetic and real data. In this section, score$^{(1)}$ is used to denoted the score function defined in (5) with the Hilbert norm associated with the Laplace Kernel. Similarly, score$^{(2)}$ is the score function defined in (6) with the data-based norm.

A. Clustering

To capture the properties of the proposed selection criteria in clustered data, we implemented the algorithm on synthetic clustered data in Figure 1. In this setup, 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$^{(1)}$ to be the score function defined in (5) with the Hilbert norm associated with the Laplace Kernel. Similarly, score$^{(2)}$ is the score function defined in (6) with the data-based norm.

The selection criterion based on score$^{(1)}$ 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$^{(1)}$ is proved in [11]. Alternatively, score$^{(2)}$ prioritizes labeling at least one sample from each cluster. Hence, after labeling 13 samples, the active learning algorithm based on score$^{(2)}$ has one sample in each cluster, but the active learning algorithm based on score$^{(2)}$ has not labeled any samples in 5 clusters.

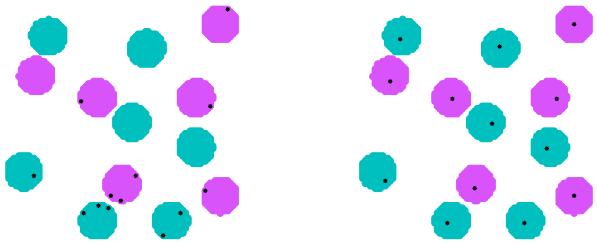

Fig. 1: Points in blue and yellow clusters are labeled $+1$ and -1, respectively. The left figure uses score$^{(1)}$ to be the score function defined in (5) with the Hilbert norm associated with the Laplace Kernel. Similarly, score$^{(2)}$ is the score function defined in (6) with the data-based norm. The first 13 samples selected by score$^{(1)}$ and score$^{(2)}$ are depicted as black dots. score$^{(2)}$ has labeled one sample from each cluster, but score$^{(1)}$ has not labeled any samples from 5 clusters. Note that score$^{(1)}$ has spent some of the sample budget to discriminate between nearby clusters with opposite labels.

B. MNIST experiments

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 $\{5, 6, 7, 8, 9\}$. We used Laplace kernel as defined in (9) with $p = 2$ and $h = 10$. In Figures 2, score$^{(1)}$ is the score function defined in (5) with the Hilbert norm associated with the Laplace Kernel. Similarly, score$^{(2)}$ is the score function defined in (6) 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 score$^{(1)}$ and score$^{(2)}$ functions on the traning 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 2 (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 2 (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.

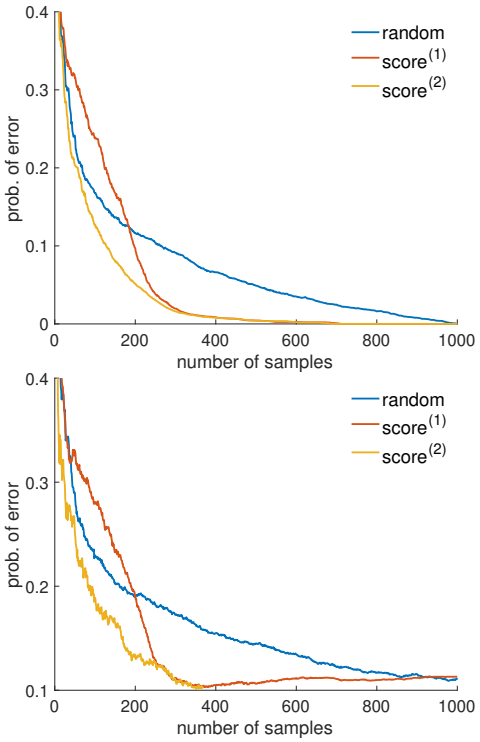

Fig. 2: Probability of error for learning a classification task on MNIST data set. The performance of three selection criteria for labeling the samples: random selection, active selection based on score$^{(1)}$, and active selection based on score$^{(2)}$. The first curve depicts the probability of error on the training set and the second curve is the probability of error on the test set.

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 $\mathsf{score}^{(2)}$ in Figure 2.

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 $\mathsf{score}^{(1)}$ and $\mathsf{score}^{(2)}$. If samples of each cluster are chosen to be labeled uniformly among clusters, we would see about 10 labeled samples in each cluster. Figure 3 shows the histogram described above for two variations of the selection criteria based on $\mathsf{score}^{(1)}$ or $\mathsf{score}^{(2)}$. We observe that selecting samples based on $\mathsf{score}^{(2)}$ is much more uniform among the clusters. On the contrary, selecting samples based on $\mathsf{score}^{(1)}$ gives much less uniform samples among clusters. In the particular example given in Figure 3, 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 $\mathsf{score}^{(1)}$.

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 $\mathsf{score}^{(1)}$ is 4.1 whereas standard deviation using $\mathsf{score}^{(1)}$ is 2.7. This shows that selection criterion based on $\mathsf{score}^{(2)}$ samples more uniformly among the clusters.

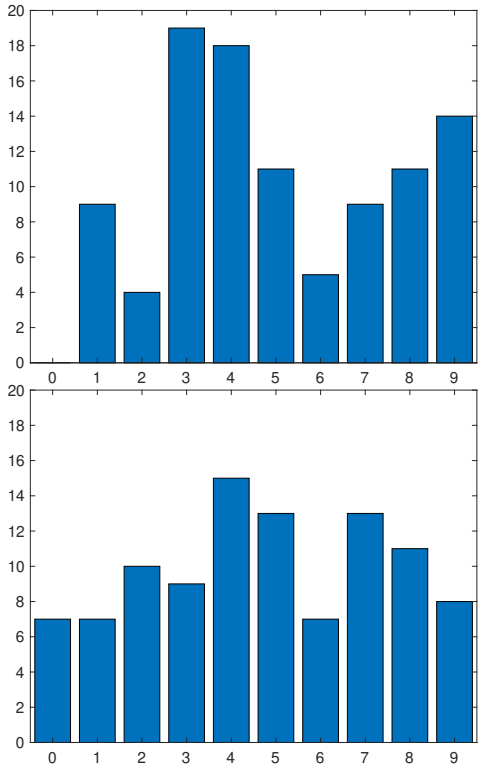

Fig. 3: The histogram of the handwritten digits associated with the labeled samples after labeling 100 samples. The first histogram is for the selection criterion $\mathsf{score}^{(1)}$ and the second histogram is for the selection criterion $\mathsf{score}^{(2)}$. Notably, $\mathsf{score}^{(1)}$ has not labeled any of the images of the digit 0.

VI. Proofs

To prove the statement of theorems presented in Section IV, we introduce some notations consistent with the notation introduced in Section III. Given a set of labeled samples $\mathcal{L} = \{(x_1,y_1),\cdots,(x_L,y_L)\}$, define the L by L matrix $\mathbf{K} = [k(x_i,x_j)]_{1\le i,j\le L}$ and vector $\mathbf{y} = [y_1,\cdots,y_L]^T$.

Recall that $\mathcal{U}$ is a set of unlabled examples. For $u \in \mathcal{U}$ and $t \in \{-1,+1\}$, let $\mathbf{a}_u = [k(x_1,u),\cdots,k(x_L,u)]^T$ and $\widetilde{\mathbf{K}}_u$ be the $L+1$ by $L+1$ matrix such that

$$\widetilde{\mathbf{K}}_u = \begin{bmatrix} \mathbf{K} & \mathbf{a}_u \\ \mathbf{a}_u^T & 1 \end{bmatrix} \quad \text{and} \quad \widetilde{\mathbf{y}}_t = \begin{bmatrix} \mathbf{y} \\ t \end{bmatrix}.$$

Let $\mathcal{B}_{d,p}(r;c)$ be the d dimensional ℓ_p ball with radius r centered at c (defined in (13)). Let $V_{d,p}(r)$ be the volume of $\mathcal{B}_{d,p}(r;0)$ with respect to the Lebesgue measure.

A. Proof of Theorem 1

The statement of theorem implies that when the data is clustered and distributed uniformly in ℓ_p balls, with centers far enough from each other, the first selected point using the score function defined in (8) is in the largest ball. To prove this, we will show that the $\mathsf{score}(c_1)$, as defined in (8) is larger than $\mathsf{score}(v)$ for any $v \notin B_1$ where c_1 is the center of B_1. Note that this does not imply that the first selected point coincides with the center of B_1. It guarantees that the largest ball contains at least one point with a score larger than that of every point in other balls.

Since $\mathcal{L} = \varnothing$, the empty set, the current interpolating function is uniformly zero everywhere $f(x) = 0$ (according to the definition (10)). According to the Equations (3) and (4), for all $u \in \mathcal{U}$, we can choose $t(u)$ to be equal to $+1$ or -1. We choose $t(u) = +1$ without loss of generality for all $u \in \mathcal{U}$.

Using (10), adding any point $u \in \mathcal{U}$ with label $t(u)$ to $\mathcal{L}$ would give the new interpolating function

$$f_u(x) := f_u^{t(u)}(x) = k(u,x) = \exp\big(-\frac{1}{h}\|x-u\|_p\big).$$

Hence, since $P_X(x)$ is uniform over $\mathcal{X} = \cup_{i=1}^M B_i$

$$\begin{aligned}\mathsf{score}(u) &= \int_{x\in\mathcal{X}} \exp\big(-\frac{2}{h}\|x-u\|_p\big)\mathrm{d}P_X(x) \\ &= \frac{1}{V}\sum_{i=1}^{M}\int_{x\in B_i}\exp\big(-\frac{2}{h}\|x-u\|_p\big)\mathrm{d}x\end{aligned}$$

where we defined $V = \sum_{i=1}^{M} V_{d,p}(r_i)$ to be the total volume of $\mathcal{X}$. So, to compute $\mathsf{score}(c_1)$,

$$\begin{aligned} V\mathsf{score}(c_1) &= \sum_{i=1}^{M} \int_{x\in B_i} \exp\big(-\frac{2}{h}\|x-c_1\|_p\big)\,\mathrm{d}x \\ &\geq \int_{x\in B_1} \exp\big(-\frac{2}{h}\|x-c_1\|_p\big)\,\mathrm{d}x \\ &= \int_{s=0}^{r_1} \exp\big(-\frac{2s}{h}\big)\,\mathrm{d}V_{d,p}(s) \end{aligned} \tag{14}$$

The integral is over a ball of radius r_1 and the integrand only depends on $\|x-c_1\|_p$ the distance from the center of the ball. Hence, we used the change of variable $s = \|x-c_1\|_p$ in the last line and used the notation defined above: $V_{d,p}(r)$ is the volume of $\mathcal{B}_{d,p}(r;0)$, a d dimensional ℓ_p ball with radius r.

For $v \notin B_1$, we want to show that $\mathsf{score}(v) \leq \mathsf{score}(c_1)$. Let $v \in B_j$ such that $j \neq 1$.

$$\begin{aligned} V\mathsf{score}(v) = &\int_{x\in B_j} \exp\big(-\frac{2}{h}\|x-v\|_p\big)\mathrm{d}x \\ &+ \sum_{i=1,i\neq j}^{M} \int_{x\in B_i} \exp\big(-\frac{2}{h}\|x-v\|_p\big)\mathrm{d}x \end{aligned} \tag{15}$$

We will bound each of above terms separately.

For any $i \neq j$ and $x \in B_i$ application of triangle inequality gives

$$\|x-v\|_p \geq \|c_i-c_j\| - \|x-c_i\| - \|v-c_j\| \geq D$$

since $v \in B_j, x \in B_i$ and $\|c_i-c_j\|_p \geq D+2r_1$, $\|x-c_i\|_p \leq r_i \leq r_1$ and $\|v-c_j\|_p \leq r_j \leq r_1$. Hence,

$$\begin{aligned} &\sum_{i=1,i\neq j}^{M} \int_{x\in B_i} \exp\big(-\frac{2}{h}\|x-v\|_p\big)\mathrm{d}x \\ &\qquad \leq \sum_{i=1,i\neq j}^{M} \int_{x\in B_i} \exp\big(-\frac{2}{h}D\big)\mathrm{d}x \\ &\qquad \leq \exp\big(-\frac{2}{h}D\big) \sum_{i=1,i\neq j}^{M} V_i\,. \end{aligned} \tag{16}$$

One can show that shows that the first term in (15) is largest when v coincides with c_j. We omit the proof of this statement for the sake of brevity, but it is presented in the longer version of the paper. Hence,

$$\begin{aligned} \int_{x\in B_j} \exp\big(-\frac{2}{h}\|x-v\|_p\big)\mathrm{d}x &\leq \int_{x\in B_j} \exp\big(-\frac{2}{h}\|x-c_j\|_p\big)\mathrm{d}x \\ &= \int_{s=0}^{r_j} \exp\big(-\frac{2s}{h}\big)\,\mathrm{d}V_{d,p}(s)\,. \end{aligned} \tag{17}$$

Equations (14), (15), (16), and (17) give

$$\begin{aligned} V\mathsf{score}(c_1) - V\mathsf{score}(v) &\geq \int_{s=r_j}^{r_1} \exp\big(-\frac{2s}{h}\big)\,\mathrm{d}V_{d,p}(s) \\ &\quad - \exp\big(-\frac{2}{h}D\big) \sum_{i=1,i\neq j}^{M} V_i \\ &\geq \exp\big(-\frac{2r_1}{h}\big)\,\big[V_1 - V_j\big] - M\,V_1 \exp\big(-\frac{2D}{h}\big)\,. \end{aligned}$$

Hence,

$$\begin{aligned} &\frac{V}{V_1}\Big[\mathsf{score}(c_1) - \mathsf{score}(v)\Big] \\ &\geq \exp\big(-\frac{2r_1}{h}\big)\,\big[1 - \frac{V_j}{V_1}\big] - M\,\exp\big(-2\frac{D}{h}\big) \\ &\overset{(a)}{\geq} \exp\big(-\frac{2r_1}{h}\big)\,\big[1 - \big(\frac{r_2}{r_1}\big)^d\big] - M\,\exp\big(-2\frac{D}{h}\big) \\ &\geq 0\,, \end{aligned}$$

where inequality (a) is due to the property that

$$V_{d,p}(r) = \frac{\big[2r\,\Gamma(1+1/p)\big]^d}{\Gamma(1+d/p)}$$

and $r_j \leq r_2$ for all $j \neq 1$. Also, the assumption

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

and $r_1 \leq h/2$ made in the statement of the theorem, yields the last inequality.

B. Proof of Theorem 2

The statement of theorem shows that if the data is clustered, and few of the clusters has been labeled so far, the algorithm selects a sample from a cluster which has not been labeled so far. To do so, without loss of generality, we show that for any $u \in B_L$, and there exists a $v \in B_{L+1}$ such that

$$\mathsf{score}(v) > \mathsf{score}(u)\,.$$

The same argument shows that for any $i \leq L$ and any $u \in B_i$, there exists a $v \in B_{L+1}$ such that $\mathsf{score}(v) > \mathsf{score}(u)$. This proves that the score of any point in the labeled balls so far is smaller than at least one point in the unlabeled clusters and hence the next point to be selected is in one of currently unlabeled balls.

We will show that for any $u \in B_L$, and there exists a $v \in B_{L+1}$ such that $\mathsf{score}(v) > \mathsf{score}(u)$. In particular, for any fixed $u \in B_1$, we choose

$$v = c_{L+1} + (u - c_1)\,. \tag{18}$$

We break the rest of the proof into five steps.

Step 1: First, we will look into the interpolator function $f(x)$ such that $f(x_i) = y_i$ for $(x_i, y_i) \in \mathcal{L}$, defined in (10).

Since $x_i \in B_i$ for $i = 1, \cdots, L$, and $\|c_i - c_j\|_p > D + 2r$, we have $\|x_i - x_j\|_p \geq D$ and $k(x_i, x_j) \leq e^{-D/h}$. Hence, matrix $\mathbf{K}$ can be decomposed as

$$\mathbf{K} = \mathbf{I}_L + e^{-D/h}\mathbf{E}$$

where $\mathbf{I}_L$ is the identity $L \times L$ matrix and matrix $\mathbf{E} = [E_{i,j}]_{1 \le i,j \le L}$ satisfies $0 \le E_{i,j} \le 1$. Hence, using Taylor series,

$$\begin{aligned}\mathbf{K}^{-1} &= \mathbf{I}_L + \sum_{n=1}^{\infty} (-1)^n e^{-nD/h} \mathbf{E}^n \\ &\overset{(a)}{=} \mathbf{I}_L + \widetilde{\mathbf{E}}^{(1)} \sum_{n=1}^{\infty} e^{-nD/h} L^{n-1} \\ &\overset{(b)}{=} \mathbf{I}_L + \frac{e^{-D/h}}{1 - Le^{-D/h}} \widetilde{\mathbf{E}}^{(1)} \overset{(c)}{=} \mathbf{I}_L + 2\, e^{-D/h} \widetilde{\mathbf{E}}^{(2)} \end{aligned} \quad (19)$$

The matrices $\widetilde{\mathbf{E}}^{(1)} = [\widetilde{E}_{i,j}]_{1 \le i,j \le L}$ and $\widetilde{\mathbf{E}}^{(2)}$ also satisfy $|\widetilde{E}^{(1)}|_{i,j} \le 1$ and $|\widetilde{E}^{(2)}|_{i,j} \le 1$. For any $n \ge 1$, the matrix $\mathbf{E}^n$ has elements smaller than L^{n-1} (This can be proved using induction over n). This gives (a). (b) is the summation of a geometric series (which holds since $D > h \log L$). (c) is due to the assumption $D > h \ln(2L)$. Plugging this into (10) gives

$$f(x) = \sum_{i=1}^{L} (y_i + \epsilon^{(f)} \gamma_i) k(x_i, x)$$

where $\epsilon^{(f)} = 2L\, e^{-D/h}$. To make the notation easier, from now on, we will use the variables γ_i with possibly different values in each line. Note that the values of γ_i depend on the elements of matrix $\widetilde{E}^{(2)}$ and realization of y_i for $i = 1, \cdots, L$. But we always have $|\gamma_i| \le 1$.

Step 2: For any $v \in B_{L+1}$, we have $\|v - x_i\|_p \ge D$ for all $i = 1, \cdots, L$. Hence, the matrix $\widetilde{\mathbf{K}}_v$ defined in (12) takes the form

$$\widetilde{\mathbf{K}}_v = \mathbf{I}_{L+1} + e^{-D/h} \mathbf{E}$$

where matrix $\mathbf{E} = [E_{i,j}]_{1 \le i,j \le L+1}$ satisfies $|E_{i,j}| \le 1$. Similar analysis as in step 1 and (19) shows that for $v \in B_{L+1}$ and any $t \in \{-1, +1\}$, (using definition of $f_v^{t(v)}(x)$ in (3)) we have

$$f_v^t(x) = \sum_{i=1}^{L} [y_i + \epsilon^{(v)} \gamma_i]\, k(x_i, x) + [t + \epsilon^{(v)} \gamma_{L+1}]\, k(v, x)$$

where $\epsilon^{(v)} = 2(L+1)\, e^{-D/h}$. Hence,

$$\begin{aligned} f_v^t(x) - f(x) &= t\, k(x, v) \\ &\quad + \big(\epsilon^{(v)} + \epsilon^{(f)}\big) \Big[\sum_{i=1}^{L} \gamma_i k(x, x_i) + \gamma_{L+1} k(x, v) \Big] \end{aligned}$$

Note that the value of the variables γ_i above might be different from the previous lines, but there exists parameters γ_i that satisfy the above equality and $|\gamma_i| \le 1$.

Step 3: For any $u \in B_L$, we will show that, $y_L f(u) \ge 0$. According to Proposition 1 in [11], discussed above in Section II, this proves that $t(u) = y_L$: our estimation of label of any sample in ball B_L is y_L, the label of the only currently labeled sample in B_L.

$$\begin{aligned} y_L f(u) &= y_1 \sum_{i=1}^{L} (y_i + \epsilon^{(f)} \gamma_i) k(x_i, u) \\ &= (1 + \epsilon^{(f)} y_L \gamma_L) k(x_L, u) + y_L \sum_{i=1}^{L-1} (y_i + \epsilon^{(f)} \gamma_i) k(x_i, u) \\ &\overset{(a)}{\ge} (1 - \epsilon^{(f)}) e^{-2r/h} - L(1 + \epsilon^{(f)}) e^{-D/h} \overset{(b)}{\ge} 0\,, \end{aligned}$$

where (a) is due to the following facts: since $x_L \in B_L$ and $u \in B_L$, we have $\|x_L - u\| \le 2r$ and $k(x_L, u) \ge e^{-2r/h}$. Also, since $u \in B_L$, for $i \le L-1$ we have $\|x_i - u\| \ge D$ and $k(x_i, u) \le e^{-D/h}$. The assumptions $D > 12h \log(2M)$, $L < M$ and the definition of $\epsilon^{(f)} = 2Le^{-D/h}$ give $\epsilon^{(f)} \le 1/100$. Then using the assumption $r < h/3$ gives (b).

Step 4: Fix $u \in B_L$ and define $d := \|u - x_L\| \le 2r$. Step 3 above proves $t(u) = y_L$. We will partition the matrix $\widetilde{\mathbf{K}}_u$ defined in (12) into the blocks corresponding to $\{x_1, \ldots, x_{L-1}\}$ and $\{x_L, x_u\}$,

$$\widetilde{\mathbf{K}}_u = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{bmatrix}$$

where $\mathbf{A}$ is a symmetric $L-1$ by $L-1$ matrix and $\mathbf{D}$ is a symmetric 2 by 2 matrix. The proof essentially follows from the fact that the elements of $\mathbf{B}$ are $k(x_i, x_L)$ and $k(x_i, x_U)$ for $i = 1, \ldots, L-1$, and hence smaller than $e^{-D/h}$. By carefully bounding the off-diagonal elements of the matrix $\widetilde{\mathbf{K}}_u^{-1}$ using Schur complements, using properties of matrix $\mathbf{D}$ and careful application of triangle inequality we show that[1] there exist parameters γ_i such that $|\gamma_i| \le 1$ and the interpolating function $f_u^{t(u)}(x)$ defined in (3) takes the form

$$\begin{aligned} f_u^{t(u)}(x) &= \Big[\frac{y_L}{1 + e^{-d/h}} + L\epsilon^{(u)} \gamma_{L+1} \Big] k(x, u) \\ &\quad + \Big[\frac{y_L}{1 + e^{-d/h}} + L\epsilon^{(u)} \gamma_1 \Big] k(x, x_L) \\ &\quad + \sum_{i=1}^{L-1} (y_i + \epsilon^{(u)} \gamma_i) K(x, x_i) \end{aligned}$$

where $\epsilon^{(u)} = 4L^3 e^{-D/h}$. Hence,

$$\begin{aligned} f_u^{t(u)}(x) - f(x) &= \frac{y_L}{1 + e^{-d/h}} k(x, u) - \frac{y_L e^{-d/h}}{1 + e^{-d/h}} k(x, x_L) \\ &\quad + \big[L\epsilon^{(u)} + \epsilon^{(f)} \big] \Big[\sum_{i=1}^{L} \gamma_i k(x, x_i) + \gamma_{L+1} k(x, u) \Big] \end{aligned}$$

Step 5: Hence, using the fact that $k(x, x') \le 1$, we get

$$\begin{aligned} &|f_v^t(x) - f(x)|^2 - |f_u^{t(u)}(x) - f(x)|^2 \\ &\quad \ge k^2(x, v) - 2(L+1)^2 \big[L\epsilon^{(u)} + \epsilon^{(v)} + 2\epsilon^{(f)} \big] \\ &\quad\quad - \frac{1}{(1 + e^{-d/h})^2} \Big[k(x, u) - e^{-d/h} k(x, x_L) \Big]^2 \end{aligned}$$

[1]We omit the proof of this statement for the sake of brevity, but it is presented in the longer version of the paper.

Since $P_X(x)$ is uniform over $\cup_{j=1}^{M} B_j$, we want to show that

$$\sum_{j=1}^{M} \int_{x\in B_j} |f_v^t(x) - f(x)|^2 - |f_u^{t(u)}(x) - f(x)|^2 \,\mathrm{d}x \geq 0\,. \tag{20}$$

To do so, we will bound the above term by

$$\begin{aligned}
&\int_{x\in B_{L+1}} k^2(x,v)\,\mathrm{d}x - 2(L+1)^2\big[L\epsilon^{(u)} + \epsilon^{(v)} + 2\epsilon^{(f)}\big]\sum_{i=1}^{M} V_i \\
&- \sum_{j=1}^{M}\int_{x\in B_j} \frac{\Big[k(x,u) - e^{-d/h}k(x,x_L)\Big]^2}{(1+e^{-d/h})^2}\,\mathrm{d}x \\
&\geq \int_{x\in B_{L+1}} k^2(x,v)\,\mathrm{d}x - \int_{x\in B_1} \frac{\Big[k(x,u) - e^{-d/h}k(x,x_L)\Big]^2}{(1+e^{-d/h})^2}\,\mathrm{d}x \\
&\quad - \Big\{2(L+1)^2\big[L\epsilon^{(u)} + \epsilon^{(v)} + 2\epsilon^{(f)}\big] + e^{-2D/h}\Big\}\sum_{i=1}^{M} V_i
\end{aligned} \tag{21}$$

where the last inequality holds since for $j \neq 1$ and $x \in B_j$, we have $k(x,u), k(x,x_L) \leq e^{-D/h}$.

Note that in (18) we defined $v = c_{L+1} + (u - c_L)$. This gives

$$\int_{B_{L+1}} k^2(x,v)dx = \int_{B_L} k^2(x,u)dx\,.$$

Hence,

$$\begin{aligned}
&\int_{x\in B_{L+1}} (1+e^{-d/h})^2 k^2(x,v)\,\mathrm{d}x - \int_{x\in B_L}\Big[k(x,u) - e^{-d/h}k(x,x_L)\Big]^2\mathrm{d}x \\
&= \int_{B_L}\Big[(1+e^{-d/h})^2 k^2(x,u) - k^2(x,u) - e^{-2d/h}k^2(x,x_L) \\
&\qquad + 2e^{-d/h}k(x,u)k(x,x_L)\Big]dx \\
&= e^{-d/h}\int_{B_L}\Big[(2+e^{-d/h})k^2(x,u) - e^{-d/h}k^2(x,x_L) \\
&\qquad + 2k(x,u)k(x,x_L)\Big]dx \\
&\overset{(a)}{\geq} e^{-2d/h}\int_{B_L} k^2(x,x_L)\Big[1 + e^{-d/h}(2+e^{-d/h})\Big]dx \\
&\overset{(b)}{\geq} e^{-4r/h}\int_{B_L} k^2(x,u)dx \overset{(c)}{\geq} e^{-6r/h}V_L \overset{(d)}{\geq} \frac{1}{10}V_L\,.
\end{aligned}$$

We defined $d = \|u - x\|_p$. This implies $k(x,u) \geq k(x,x_L)e^{-d/h}$ which gives (a). (b) uses $d \leq 2r$. For $x \in B_L$, we have $\|u - x\|_p \leq 2r$. This gives inequality (c). The assumption $\frac{h}{3}$ implies $r < \frac{h}{6}\ln 10$ which gives (d).

The assumption $D \geq 12h\ln(2M)$ implies $D \geq 6h\ln(2LM)$ (since $L < M$) which gives

$$2(L+1)^2\big[L\epsilon^{(u)} + \epsilon^{(v)} + 2\epsilon^{(f)}\big] + e^{-2D/h} < \frac{1}{15M}\,.$$

Plugging the above two statements in (21) gives the desired result.

So for any $u \in \cup_{i=1}^{L} B_i$, there exists a $v \in \cup_{i=L+1}^{M} B_i$ which has larger score. Hence, the selection criterion based on $\mathsf{score}^{(2)}$ would always pick a sample from a new ball to be labeled.

REFERENCES

[1] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*, 2019.

[2] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118*, 2018.

[3] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

[4] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*, pages 2300–2311, 2018.

[5] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 540–548, 2018.

[6] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(Nov):2399–2434, 2006.

[7] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? *arXiv preprint arXiv:1806.09471*, 2018.

[8] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017.

[10] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

[11] Mina Karzand and Robert D Nowak. Active learning in the overparameterized and interpolating regime. *arXiv preprint arXiv:1905.12782*, 2019.

[12] Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel" ridgeless" regression can generalize. *arXiv preprint arXiv:1808.00387*, 2018.

[13] Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pages 3331–3340, 2018.

[14] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[15] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[16] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[17] Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*, 2017.

[18] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 824–831. ACM, 2005.

[19] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

[20] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2017.