Geometric Analysis of Uncertainty Sampling for Dense Neural Network Layer

Aziz Koçanaoğulları, Niklas Smedemark-Margulies, Murat Akcakaya, and Deniz Erdoğmuş

***Abstract*—For model adaptation of fully connected neural network layers, we provide an information geometric and sample behavioral active learning uncertainty sampling objective analysis. We identify conditions under which several uncertainty-based methods have the same performance and show that such conditions are more likely to appear in the early stages of learning. We define riskier samples for adaptation, and demonstrate that, as the set of labeled samples increases, margin-based sampling outperforms other uncertainty sampling methods by preferentially selecting these risky samples. We support our derivations and illustrations with experiments using Meta-Dataset, a benchmark for few-shot learning. We compare uncertainty-based active learning objectives using features produced by SimpleCNAPS (a state-of-the-art few-shot classifier) as input for a fully-connected adaptation layer. Our results indicate that margin-based uncertainty sampling achieves similar performance as other uncertainty based sampling methods with fewer labelled samples as discussed in the novel geometric analysis.**

***Index Terms*—Active learning, few-shot learning, information geometry, margin sampling, uncertainty sampling.**

I. Introduction

RECENTLY, deep neural networks have been commonly used for hypothesis learning in the context of various regression and classification problems. These models require large labeled data sets to achieve good generalization performance [1]. Obtaining large labeled data sets is costly; several approaches exist to overcome this limitation. For example, model adaptation with few samples (informally referred to as zero-, one-, few-shot learning) may enable transformation of a hypothesis model (e.g., a classifier) trained for one specific task to a hypothesis model suitable for another task through minimal changes to its structure and parameters [1], [2]. Model adaptation for deep neural networks between classification tasks is typically achieved by adjusting the last few layers [3], [4]. Active learning is crucial in cases where rapid adaptation is required with limited data labeling resources [5].

Related Work: As described in Settle's work [6], active learning objectives often combine the following measures: uncertainty in probability space to select ambiguous samples [7], density in data space (e.g. N nearest neighbors [8]), expected model change (e.g. absolute sum of the gradients [9]), and expected error reduction (e.g. maximizing mutual information [10]). Other proposed methods include influence functions [11], and representer point based selection [12]. In the existing active learning literature, uncertainty sampling methodologies (Entropy Sampling (ES) [13], Confidence Sampling (CS) [14], and Margin Sampling (M) [15]) are often used as baseline comparisons [16], [17] due to their low computational overhead. Existing work usually reports the best performing uncertainty sampling method based on the test performance results, but performance comparison across (ES), (CS) and (M) and justification of the performance differences of these uncertainty methods across different datasets are omitted. The fundamental differences across (ES), (CS) and (M) are studied in the literature with extensive experimentation in different domains [18], [19]. Uncertainty sampling has also been comprehensively studied in parameter estimation for logistic regression models [20]. We suggest that the performance differences across (ES), (CS) and (M) across different testing datasets occur due to the sampling behaviour differences in these uncertainty sampling methods during actively updating the models. However, none of the existing work provide an explanation to these differences. Here, we aim to explain the sampling behavior differences for uncertainty based methods from an information geometric perspective, specifically through the geometrical analysis of the unlabeled sample predictions. We consider a fully connected neural network layer and demonstrate that by design (M) performs better in selecting riskier samples (i.e., the samples that are geometrically located in highly uncertain locations in the probability simplex) that enables it to achieve similar performances to (ES) and (CS) with fewer samples.

Contributions: As mentioned above, uncertainty sampling methods have been commonly used and they were compared among each other and against other methods always through their performances on different datasets [19], [21]. We propose here a novel analytical approach to compare these methods, such an analytical approach does not exist in the literature. Specifically, (i) We identify conditions under which uncertainty-based methods have equal performance. (ii) We use information geometry to demonstrate the behavior of sample selection for (ES), (CS) and (M) and highlight that (M) selects samples from highly uncertain locations in the probability

Manuscript received October 28, 2020; accepted April 4, 2021. Date of publication April 9, 2021; date of current version May 11, 2021. This work was supported by the NIH under Grant R01DC009834; by the DARPA under Grant SC1821301; and by the NSF under Grant CNS-1544895, Grant IIS-1715858, Grant IIS-1717654, Grant IIS-1844885, and Grant IIS-1915083. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Z. Jane Wang. *(Corresponding author: Aziz Koçanaoğulları.)*

Aziz Koçanaoğulları and Deniz Erdoğmuş are with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: akocanaogullari@ece.neu.edu; erdogmus@ece.neu.edu).

Niklas Smedemark-Margulies is with the Khoury College of Computer Science, Northeastern University, Boston, MA 02115 USA (e-mail: smedemark-margulie.n@northeastern.edu).

Murat Akcakaya is with the Department of Electrical and Computer Engineering, Pittsburgh University, Pittsburgh, PA 15261 USA (e-mail: akcakaya@pitt.edu).

Digital Object Identifier 10.1109/LSP.2021.3072292

simplex. (iii) We validate our analysis in a few-shot learning scenario.

Preliminaries: Let $(\mathcal{X}, \mathcal{Y})$ denote the domain of data and labels over C different classes $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_C\}$. Let $(x \in \mathbb{R}^d, y)$ be a single data example and its corresponding one-hot label vector. To estimate the class label, we fit a parametric model $h_\theta : \mathcal{X} \to \Delta_C$ where Δ_C is the C-dimensional simplex:

$$\Delta_C = \{(p_1, p_2, \ldots, p_C) \in \mathbb{R}^C | p_i > 0 \; \forall i, \sum_i p_i = 1\}$$

Model parameters θ are fit by minimizing a loss function $L : (h_\theta(\mathcal{X}), \mathcal{Y}) \to \mathbb{R}$, the model parameters are optimized appropriately. In this paper, we specifically focus on the following parameterized model;

$$\begin{aligned} h_\theta(x) &= \text{softmax}\left(\theta \begin{bmatrix} x \\ 1 \end{bmatrix}\right) \text{ where } \theta \in \mathbb{R}^{C \times (d+1)} \\ \text{softmax}(a)_i &= \frac{e^{a_i}}{\sum_{j=1}^d e^{a_j}} \text{ where } a \in \mathbb{R}^d \end{aligned} \tag{1}$$

Here θ is the parameter matrix and softmax : $\mathbb{R}^n \to \Delta_n$ denotes the operator that maps the linearly transformed x to the probability simplex [22]. The hypothesis class in (1) is a linear logistic regression function, i.e., a fully connected neural network layer followed by softmax nonlinearity.

In many applications, hypothesis learning can be achieved using a combination of labelled and unlabelled data s.t. $(\mathcal{X}, \mathcal{Y}) = (\{\mathcal{X}_L, \mathcal{X}_U\}, \{\mathcal{Y}_L, \mathcal{Y}_U\})$ where L and U denote subsets of labelled and unlabelled data respectively; Y_U is not available to the learner. We focus on the empirical risk minimization framework, in which optimal parameters for the model h are computed as:

$$\hat{\theta}_L = \arg\min_\theta \frac{1}{|\mathcal{X}_L|} \sum_{(x,y) \in (\mathcal{X}_L, \mathcal{Y}_L)} L(h_\theta(x), y)$$

For the training of the last fully connected layer of a classification neural network, minimization of average cross entropy loss is considered $L(h_\theta(x), y) = H(h_\theta(x), y)$ [23]:

$$\begin{aligned} &\arg\min_\theta H(\vec{y}, h_\theta(x)) = \arg\min_\theta - \textstyle\sum_i y_i \log([h_\theta(x)]_i) \\ &= \arg\min_\theta - \log([h_\theta(x)]_{\hat{i}}) \text{ where } \hat{i} = \arg\max_j y_j \\ &= \arg\max_\theta [h_\theta(x)]_{\hat{i}} \text{ where } \hat{i} = \arg\max_j y_j \end{aligned} \tag{2}$$

Actively learning a model includes an agent that selects *anchor* samples $x_a \in \mathcal{X}_U$ according to a sampling objective f, and receives the corresponding label y_a from an oracle to update the model:

$$\begin{aligned} x_a &= \arg\max_{x \in \mathcal{X}_U} f(x, \mathcal{X}_L, \mathcal{Y}_L, h_\theta) \\ \hat{\theta}_{L \cup a} &= \arg\min_\theta \frac{1}{|\mathcal{X}_L|} \sum_{(x,y) \in (\mathcal{X}_L \cup x_a, \mathcal{Y}_L \cup y_a)} H(\vec{y}, h_\theta(x)) \\ (\mathcal{X}_L, \mathcal{Y}_L) &\leftarrow (\mathcal{X}_L, \mathcal{Y}_L) \cup (x_a, y_a) \\ (\mathcal{X}_U, \mathcal{Y}_U) &\leftarrow (\mathcal{X}_U, \mathcal{Y}_U) \setminus (x_a, y_a) \end{aligned} \tag{3}$$

In (3), f is designed to decrease the loss value as fast as possible by selecting meaningful samples. For f, we consider (ES), (CS) and (M) objectives presented in Table I and the geometry of each method based on their objectives is illustrated in Fig. 1.

II. Analysis

The performance differences among (ES), (CS) and (M) selection objectives arise from the sequence of anchor samples x_a selected. Through Proposition 1, we show that these objectives have the same performance for 2-class classification. Further analyses then discuss the performance differences among the methods.

TABLE I
UNCERTAINTY SAMPLING METHODS THAT FORM THE BASIS OF ACTIVE LEARNING METHODOLOGIES

Identifier	Root	Selection Method	Ref
(R)	random	$x_a = \text{random}(\mathcal{X}_U)$	–
(ES)	entropy	$x_a = \arg\max_{x \in \mathcal{X}_U} - \sum_i [h_\theta(x)]_i \log[h_\theta(x)]_i$	[13]
(CS)	confidence	$x_a = \arg\min_{x \in \mathcal{X}_U} [h_\theta(x)]_i$ s.t. $i = \arg\max_c [h_\theta(x)]_c$	[14]
(M)	margin	$x_a = \arg\min_{x \in \mathcal{X}_U} [h_\theta(x)]_i - [h_\theta(x)]_j$ s.t. $i = \arg\max_c [h_\theta(x)]_c$, $j = \arg\max_{c \neq i} [h_\theta(x)]_c$	[15]

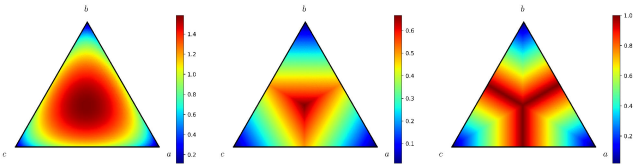

Fig. 1. Geometry of the given methods in Δ_3. Figures from left to right represent the values of the objective functions presented in Table I (ES), (CS) and (M) respectively.

Proposition 1: Given 2 class case $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2\}$ and the set $\mathcal{X}$ with $h_\theta(x) = p \in \Delta_2$ then;

$$\begin{aligned} x_a &= \arg\max_{p \in h_\theta(\mathcal{X})} H(p) \text{ (ES)} \\ &= \arg\min_p \max_i p_i \text{ (CS)} \\ &= \arg\min_p \max_{j \neq i} \max_i p_i - p_j \text{(M)} \end{aligned}$$

Proof: Let us denote the anchor sampling methods; (i) $\arg\max_p H(p)$, (ii) $\arg\min_p \max_i p_i$, (iii) $\arg\min_p \max_{j \neq i} \max_i p_i - p_j$. $p \in \Delta_2 \Rightarrow p_i = 1 - p_{\neq i}$.

$$\arg\min_p \max_{j \neq i} \max_i p_i - p_j = \arg\min_p \max_i p_i - p_{\neq i}$$

$$= \arg\min_p \max_i 2p_i - 1 = \arg\min_p \max_i p_i \Rightarrow \text{(ii)} \equiv \text{(iii)}$$

Similarly,

$$\begin{aligned} \arg\max_p H(p) &= \arg\max_p - \sum_i p_i \log(p_i) \\ &= \arg\max_p - p_i \log(p_i) - p_{\neq i} \log(p_{\neq i}) \\ &= \arg\max_p - p_i \log(p_i) - (1 - p_i) \log(1 - p_i) \end{aligned}$$

WLOG assume $p_i \geq p_{\neq i}$ then, $1 \geq p_i \geq 0.5$ and $-p_i \log(p_i) - (1 - p_i) \log(1 - p_i)$ is monotonically decreasing wrt. $p_i \to$ given $p, q \in \Delta_2$ $\max_i p_i > \max_i q_i \Rightarrow H(p) < H(q) \Rightarrow \arg\max_p H(p) = \arg\min_p \max_i p_i \Rightarrow$ (i) $\equiv$ (ii). Then (i) $\equiv$ (ii) $\equiv$ (iii). ■

In a C-class classification problem, model h_θ makes a correct decision for a sample and label tuple $(x, y) \in (\mathcal{X}, \mathcal{Y})$ if $\arg\max_i y_i = \arg\max_i [h_\theta(x)]_i$. Therefore $\forall h_\theta(x) \in \Delta_n$ the critical boundary for a class i is formed where coordinate i is tied with a single other coordinate j. In other words, $\exists j$ s.t. $[h_\theta(x)]_i = [h_\theta(x)]_j \geq [h_\theta(x)]_{k \neq i,j} \; \forall k$.

TABLE II

A FULLY CONNECTED LAYER ON TOP OF THE BACKBONE NETWORK ARCHITECTURE IS USED. THE AVERAGE NUMBER OF TRAINING SAMPLES ARE ≈ 355 WHERE NUMBER OF TEST SAMPLES ARE ≈ 90. IN THE TABLE *ROOT-ACC* AND *MAX-ACC* REFER TO ACCURACY ACHIEVED IN TEST-SET IF ALL TRAINING-SET IS USED AND MAXIMUM ACCURACY ACHIEVED IN ANY POINT OF ACTIVE LEARNING RESPECTIVELY. WE REPORT THE RATIO OF TRAINING SAMPLES USED TO CARDINALITY OF THE ENTIRE TRAINING SET TO REACH A NEIGHBORHOOD OF *ROOT-ACCURACY*. THEREFORE, A SMALLER NUMBER REPRESENTS FEWER SAMPLES USED AND HENCE THE LESS THE BETTER. IT IS APPARENT THAT **MARGIN** SAMPLING (M) OUTPERFORMS OTHER SELECTION METHODS AND ACHIEVES THE CONFIDENCE RANGE USING LESS PERCENTAGE OF THE TRAINING SET LABELLED.

		root-acc -15%				root-acc -10%				root-acc -5%				root-acc -1%			
data-set	root-acc (%)	(R)	(ES)	(CS)	(M)	(R)	(ES)	(CS)	(M)	(R)	(ES)	(CS)	(M)	(R)	(ES)	(CS)	(M)
aircraft	83	0.24±0.17	0.29±0.20	0.24±0.17	**0.23±0.17**	0.29±0.17	0.40±0.22	0.27±0.17	**0.26±0.17**	0.43±0.20	0.62±0.22	0.38±0.18	**0.33±0.17**	0.64±0.24	0.80±0.18	0.56±0.20	**0.47±0.20**
cifar-10	78	0.24±0.16	0.37±0.22	0.21±0.16	**0.24±0.15**	0.31±0.16	0.49±0.23	0.25±0.16	**0.24±0.15**	0.47±0.20	0.70±0.19	0.36±0.17	**0.33±0.16**	0.68±0.22	0.85±0.14	0.52±0.21	**0.47±0.21**
cu-birds	76	0.24±0.17	0.29±0.20	0.24±0.17	**0.23±0.17**	0.29±0.17	0.40±0.22	0.27±0.17	**0.26±0.17**	0.43±0.19	0.62±0.22	0.38±0.18	**0.33±0.16**	0.64±0.24	0.80±0.17	0.56±0.20	**0.48±0.20**
fungi	40	0.50±0.14	0.50±0.14	0.50±0.14	0.50±0.14	0.50±0.14	0.50±0.14	0.50±0.14	0.50±0.14	0.53±0.15	0.54±0.17	0.52±0.14	**0.51±0.14**	0.62±0.20	0.65±0.21	**0.57±0.16**	0.58±0.16
ms-coco	48	0.38±0.12	0.39±0.12	**0.38 ±0.11**	0.39±0.12	0.40±0.12	0.44±0.15	**0.39±0.12**	**0.39±0.12**	0.52±0.17	0.58±0.20	0.45±0.13	**0.44±0.13**	0.73±0.20	0.76±0.22	0.62±0.18	**0.58±0.17**
omniglot	91	0.76±0.15	0.76±0.15	0.76 ±0.15	0.76±0.15	0.76±0.15	0.76±0.15	0.76±0.15	0.76±0.15	0.76±0.15	0.76±0.15	0.76±0.15	0.76±0.15	0.83±0.11	0.83±0.11	0.80±0.12	**0.78±0.12**
quickdraw	76	0.44±0.10	0.44±0.11	0.44 ±0.11	0.44±0.11	0.44±0.10	0.46±0.12	0.44±0.10	0.44±0.10	0.52±0.13	0.60±0.17	0.50±0.11	**0.48±0.11**	0.78±0.17	0.84±0.15	0.66±0.14	**0.63±0.14**
traffic-sign	77	0.39±0.12	0.39±0.13	**0.38 ±0.12**	0.39±0.12	0.39±0.12	0.41±0.16	0.39±0.12	0.39±0.12	0.42±0.14	0.50±0.24	0.39±0.12	0.39±0.12	0.50±0.19	0.63±0.30	0.41±0.13	**0.40±0.12**
vgg-flower	91	0.29±0.21	0.30±0.21	0.29 ±0.21	0.29±0.21	0.30±0.21	0.36±0.25	0.30±0.21	0.29±0.21	0.42±0.20	0.61±0.28	0.33±0.20	**0.32±0.20**	0.65±0.23	0.85±0.20	0.44±0.20	**0.41±0.19**
average	-	0.38	0.41	0.38	0.38	0.40	0.47	0.40	**0.39**	0.50	0.61	0.45	**0.43**	0.67	0.78	0.57	**0.53**

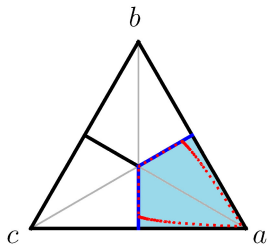

Fig. 2. A three class decision boundaries on Δ_3. The highlighted area corresponds to where the probability mass of a in a given probability vector $\in \Delta_3$ is the highest. Therefore a hypothesis resulting a probability within that area correctly classifies a. Dashed red line represent an equi-probability contour of a logistic normal distribution that passes through the uniform distribution.

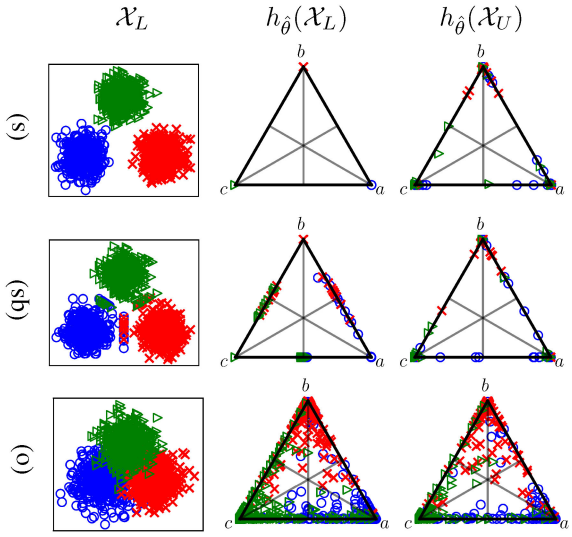

Fig. 3. Separated (s), quasi-separated (qs) and overlapping (o) labelled data $\mathcal{X}_L$ is used to determine a stationary point $h_{\hat{\theta}}$. Observe that (s), (qs) yields unlabelled samples on the lines that force exact same results in (ES), (CS) and (M). (o) on the other hand allows selection methods operate in 2D.

Assume a three class classification as illustrated in Fig. 2 where the labels are denoted with a, b, c. Then in order to correctly classify a for example, the hypothesis should result a probability vector where a has the highest probability mass (blue highlighted area). Hence all possible probability vector with a having the max probability contribute to the true selection. In the figure the decision boundary for class a is also visualized with the blue-bold lines. Given $p = h_\theta(x)$, these lines satisfy $\exists i, j$ s.t. $p_i = p_j \geq p_k \, \forall k$. In other words, the model is uncertain between two competitors. Points $x \in \mathcal{X}$ that result in $p = h_\theta(x)$ close to these decision boundaries are riskier samples.

(ES) has a frail confidence assessment: Let $\{x1, x2\} \subset \mathcal{X}_\mathcal{U}$ where $\arg\max_i y1_i = \arg\max_i y2_i = 1$ and $p = h_{\hat{\theta}}(x1), q = h_{\hat{\theta}}(x2) \in \Delta_{10}$ with $p = [.6, .4 - 8\epsilon, \epsilon, \cdots]$ where $0 < \epsilon << 1$ and $q = [.7, .0\bar{3}, \cdots]$. It is apparent that q is more confident on the true class however due to $0.97 \approx H(p) < H(q) \approx 1.82$ (ES) selects q over p. For this example (M) and (CS) captures the notion of confidence which is determined by the maximum posterior probability and selects the sample with lesser confidence.

(CS) makes decisions using single value: Let $\{x1, x2, x3, x4\} \subset \mathcal{X}_\mathcal{U}$ with $p1 = h_{\hat{\theta}}(x1), p2 = h_{\hat{\theta}}(x2) \in \Delta_{10}, q1 = h_{\hat{\theta}}(x3), q2 = h_{\hat{\theta}}(x4) \in \Delta_{10}$ with $p1 = [.5, .5 - 8\epsilon, \epsilon, \cdots]$, $p2 = [.6, .4 - 8\epsilon, \epsilon, \cdots]$ where $0 < \epsilon << 1$ and $q1 = [.5, 0.0\bar{5}, \cdots]$, $q2 = [.6, 0.0\bar{4}, \cdots]$ and we compare ps then qs. Since confidence level on a particular class differs the same amount. 1 (CS) treats p and q cases the same, for the ps the 2nd best class is still a legitimate competitor, whereas for qs there was no other competitor. However, (M) makes a distinction between these cases by incorporating another element to select the sample closer to decision boundary (p).

Sample Behavior: Ultimately, the differences in performance occur due to the different geometric locations of the data in the simplex (i.e., $h_\theta(x) \in \Delta_n$) compared to the geometric selection regions of different objectives. Let us observe the different outcome possibilities from the fully connected layer.

Consider Fig. 3 for Δ_3 for three scenarios of h_θ outcomes. Given $(x^l, y^l) \in (\mathcal{X}_L, \mathcal{Y}_L)$ and $(x^u, y^u) \in (\mathcal{X}_U, \mathcal{Y}_U)$ if $\exists \theta$ s.t.: (s) separation $(\forall (x^l, y^l) \arg\max_i [h_\theta(x^l)]_i = \arg\max_i y_i^l)$, (qs) quasi-separation $(\forall (x^l, y^l) \;\; \arg\max_i [h_\theta(x^l)]_i = \arg\max_i y_i^l$ and $\exists (x^l, y^l)$ s.t. $|\arg\max_i h_\theta(x^l)| \geq 2$. e.g. $h_\theta(x^l) = [0.5, 0.5, 0, \cdots]$, $y^l = [1, 0, \cdots]$), (o) overlap $(\exists (x^l, y^l)$ where $\arg\max_i [h_\theta(x^l)]_i \neq \arg\max_i y_i^l)$ then a stationary point $\hat{\theta} = \arg\inf_\theta (1/|\mathcal{X}_L|) \sum_{(x,y) \in (\mathcal{X}_L, \mathcal{Y}_L)} H(\vec{y}, h_\theta(x))$ only exists for (o).

We refer the reader to Albert's work [24] for the detailed proof showing (s) and (qs) project samples to a two class decision as shown in Fig. 3. In Proposition 1 we already discussed if

samples reside on a line which is the case for (s) and (qs), the sampling methodologies behave the same and hence we further investigate the case of (o). Note that it is empirically known that the differences in the performances among (CS), (ES) and (M) increase as the training progresses. As also discussed above, the performances differ only in the case of (o). With the following proposition, we show that during the training, the probability of achieving the case of (o) as the outcome of h_θ in a dataset increases as more data is incorporated into the labeled pool $\mathcal{X}_L$ for training;

Proposition 2: Let h_θ be an arbitrary model and $(\mathcal{X}, \mathcal{Y}) = \{(x, y)|x \sim f_y\}$ where f_y is an arbitrary distribution identified with the label y. Let $(\mathcal{X}_1, \mathcal{Y}_1) \subseteq (\mathcal{X}_2, \mathcal{Y}_2) \subseteq (\mathcal{X}, \mathcal{Y})$ then probability of (o) in $(\mathcal{X}_2, \mathcal{Y}_2)$ is greater or equal than probability of (o) in $(\mathcal{X}_1, \mathcal{Y}_1)$.

Proof: Let $A := \exists$ overlap $\in (\mathcal{X}_1, \mathcal{Y}_1)$, $B := \exists$overlap $\in (\mathcal{X}_2, \mathcal{Y}_2) \setminus (\mathcal{X}_1, \mathcal{Y}_1) \Rightarrow A + B := \exists$ overlap $\in (\mathcal{X}_2, \mathcal{Y}_2)$. Trivially $p(A + B) \geq p(A) + p(B) \Rightarrow p(A + B) \geq p(A)$. ■

In summary, Proposition 2 states that as the training progresses, the probability of the case (o) increases which implies that for active learning the (CS), (ES) and (M) will start to have different performances.

Special case with Gaussian data: For the case of (o), consider samples originating from multi-variate Gaussian distributions for a C-class classification s.t. $\forall(x, y) \in (\mathcal{X}, \mathcal{Y})$, $x = \mathcal{N}(\mu_y, \Sigma_y)$. Note that the equi-probability contours are not centered around the center of the Δ_C (simplex) but the corners. Since x is normally distributed, linear combinations with a scalar shift of the random variables also follow a normal distribution. Moreover, the inverse of softmax operator is well approximated as central logratio transform clr(.) [22], [25]:

$$\text{clr}(p) = \left[\frac{p_1}{g(p)}, \ldots, \frac{p_C}{g(p)}\right] \text{ where } g(p) = (\prod_i p_i)^{1/C}$$

Hence the outputs of the model defined in (1) follow a logistic-normal distribution for which the pdf is;

$$f_{\Delta_C}(p; m, s) = J_C \frac{1}{|2\pi s|^{\frac{1}{2}}} \exp(-\frac{1}{2}(p' - m)^T s^{-1}(p' - m)) \text{ where } p' = p/g(p) \tag{4}$$

As discussed in [22], it is possible to find equi-probability contours within Δ_C. WLOG assume we are interested in $\mathcal{H}_1$ with an ideal $m = [1 - \varepsilon, \varepsilon, \ldots, \varepsilon]$ with $0 < \varepsilon << 1$. In Fig. 2 we visualize a sample equi-probability contour with the dashed red line. Hence in terms of the probability geometry, the assessment of riskier samples are not centric as guided by the entropy but they follow a pattern that are analogous with the critical boundaries and hence a perspective that is centered around the corners is more preferable. Comparison between Figures 1 and 2 show that (M) captures these equi-probable regions but not (CS) or (ES).

III. Experiments and Results

Paradigm: Using a few-shot learning scenario, we compare (ES), (CS) and (M). Also we consider the random selection case, see Table I. In few-shot learning approaches, model parameters are updated starting from a checkpoint [1], [2]. In the cases where adaptation needs to be fast, especially in deep feature models, a backbone that is already trained on a large labelled dataset (e.g. image classification ResNet [27]) is kept constant and an adaptation layer is further adjusted to generalize across an unseen task. Specifically, as the backbone, in our experiments we use the model presented in [3] and later simplified in [4] which is by design learned to output features that follow Gaussian distribution. We then actively learn a fully connected layer during adaptation.

Dataset and Experiment: We use *Meta-dataset* [28], a benchmark for few-shot learning and image classification that comprises the following labelled image datasets: ILSVRC-2012 [29], Omniglot [30], FGVC-Aircraft [31], CUB-200-2011 [32], Describable Textures [33], QuickDraw [34], FGVCx Fungi, VGG Flower [35], Traffic Signs [36] and MSCOCO [37]. We follow the train-test splits provided by *Meta-dataset* in our experiments. We define '*root-acc*'; as the test set accuracy achieved using all training labels set is available. We evaluate active learning methods by comparing what percentage of the training set must be queried before achieving the *root-acc*. In our experiments, we initialize the system by providing labels for 5% of the training samples and training an initial model: $|\mathcal{X}_U| = 0.05 \times |\mathcal{X}_{U \cup L}|$. At each iteration, we select a batch of 10-samples to be labelled according to the objectives presented in Table I. At each iteration, the models are also updated to a stationary point.

Results: Results are presented in Table II. In this Table, the rows represent the results for different datasets and the average of all is presented as the final row. The table is divided into 4 grouped-columns, where each group denotes a model reaching a pre-determined performance value close to root-acc. For example, where root-acc 85%, the column root-acc-15% represents a model achieving a performance ≥ 70. Each column in each group presents the results of a different sample selection method from Table I. We present the mean and standard deviation for the percentage of the dataset used to achieve the pre-determined performance value. Lower mean values indicate that the method requires less data to match the performance of the other sampling methods. The results show that (M) picks the samples that result in faster increase in performance. Moreover, we observe that the gap between methodologies increases as labelled data size increases (to achieve higher performance more data is required) which is also stated in Prop. 2.

IV. Conclusion

In this work we provide an analysis of actively adapting a fully connected final layer in a network architecture in a model-adaption setting. Specifically, we focused on uncertainty sampling methods that are widely used as a sanity check in active learning tasks. We have shown that geometrically, fully connected layer behavior and sample positioning by fact strengthens margin sampling over other uncertainty based approaches. Empirically, we validated the claims in a few-shot learning setting where a fully connected adaptation layer exists. With that knowledge, it is possible to propose proxy gradient methods that leverage margin instead of selection based on mutual information surrogates.

Acknowledgment

The authors would like to thank Jan-Willem van de Meent for valuable input that helped them improve the paper.

REFERENCES

[1] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. the 34th Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[2] B. Oreshkin, P.R.López, and A.Lacoste, "Tadam: Task dependent adaptive metric for improved few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 721–731.

[3] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner, "Fast and flexible multi-task classification using conditional neural adaptive processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7957–7968.

[4] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal, "Improved few-shot visual classification, in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,*2019, pp. 14493–14502.

[5] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales,"Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[6] B. Settles, "Active learning literature survey," Dept. Comput. Sci. Univ. Wisconsin-Madison, Tech. Rep., 2009.

[7] D. D. Lewis and J.Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. Mach. Learn. Proc.*, 1994, pp. 148–156.

[8] C. Berlind and R. Urner, "Active nearest neighbors in changing environments," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1870–1879.

[9] Y. Yuan, S.-W. Chung, and H.-G. Kang, "Gradient-based active learning query strategy for end-to-end speech recognition," in *Proc. ICASSP IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 2832–2836.

[10] J. Sourati, M. Akcakaya, J. G. Dy, T. K. Leen, and D.Erdogmus, "Classification active learning based on mutual information," *Entropy*, vol. 18, no. 2, pp. 51–72, 2016.

[11] P. W.Koh and P.Liang, "Understanding black-box predictions via influence functions," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1885–1894.

[12] C.-K. Yeh, J. Kim, I. E.-H. Yen, and P. K. Ravikumar, "Representer point selection for explaining deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 9291–9301.

[13] C. E. Shannon, "A. mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, 1948.

[14] D. D. Lewis and W. A. Gale, "A. sequential algorithm for training text classifiers," in *Proc. SIGIR'94*, 1994, pp. 3–12.

[15] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proc. Int. Symp. Intell. Data Anal.*, 2001, pp. 309–318.

[16] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H1aIuk-RW

[17] 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, Dec. 2017.

[18] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2008, pp. 1070–1079.

[19] Y. Yang, Z. Ma, F. Nie, X. Chang, and A. G. Hauptmann, "Multi-class active learning by uncertainty sampling with diversity maximization," *Int. J. Comput. Vis.*, vol. 113, no. 2, pp. 113–127, 2015.

[20] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: An evaluation," *Mach. Learn.*, vol. 68, no. 3, pp. 235–265, 2007.

[21] M. Li and I. K. Sethi, "Confidence-based active learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1251–1261, Aug. 2006.

[22] J. Aitchison, "The statistical analysis of compositional data," *J. Roy. Stat. Soc.: Ser. B (Methodological)*, vol. 44, no. 2, pp. 139–160, 1982.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.

[24] A. Albert and J. A. Anderson, "On the existence of maximum likelihood estimates in logistic regression models," *Biomet*, vol. 71, no. 1, pp. 1–10, 1984.

[25] J. Aitchison, "Logratios and natural laws in compositional data analysis," *Math. Geol.*, vol. 31, no. 5, pp. 563–580, 1999.

[26] D. A. Cohn, Z.Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, 1996.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[28] E. Triantafillou *et al.*, "Meta-dataset: A dataset of datasets for learning to learn from few examples," in *Proc. Int. Conf. Learn. Representations*, 2019.

[29] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[30] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum,"Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[31] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," 2013. [Online]. https://www.robots.ox.ac.uk/~vgg/data/fgvc-aircraft/.

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD birds-200-2011 dataset," 2011. [Online]. Available: http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

[33] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3606–3613.

[34] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, and N. Fox-Gieg, "The Quick, Draw!-AI experiment," Mount View, CA, USA, 2016. Accessed: Feb. 17, 2018. [Online]. Available: https://quickdraw.withgoogle.com/

[35] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, 2008, pp. 722–729.

[36] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–8.

[37] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.