Neural Active Learning with Performance Guarantees

Zhilei Wang
New York University
New York, NY 10012
zhileiwang92@gmail.com

Pranjal Awasthi
Google Research
New York, NY 10011
pranjalawasthi@google.com

Christoph Dann
Google Research
New York, NY 10011
chrisdann@google.com

Ayush Sekhari
Cornell University
Ithaca, NY 14850
ayush.sekhari@gmail.com

Claudio Gentile
Google Research
New York, NY 10011
cgentile@google.com

Abstract

We investigate the problem of active learning in the streaming setting in non-parametric regimes, where the labels are stochastically generated from a class of functions on which we make no assumptions whatsoever. We rely on recently proposed Neural Tangent Kernel (NTK) approximation tools to construct a suitable neural embedding that determines the feature space the algorithm operates on and the learned model computed atop. Since the shape of the label requesting threshold is tightly related to the complexity of the function to be learned, which is a-priori unknown, we also derive a version of the algorithm which is agnostic to any prior knowledge. This algorithm relies on a regret balancing scheme to solve the resulting online model selection problem, and is computationally efficient. We prove joint guarantees on the cumulative regret and number of requested labels which depend on the complexity of the labeling function at hand. In the linear case, these guarantees recover known minimax results of the generalization error as a function of the label complexity in a standard statistical learning setting.

1 Introduction

Supervised learning is a fundamental paradigm in machine learning and is at the core of modern breakthroughs in deep learning [29]. A machine learning system trained via supervised learning requires access to labeled data collected via recruiting human experts, crowdsourcing, or running expensive experiments. Furthermore, as the complexity of current deep learning architectures grows, their requirement for labeled data increases significantly. The area of *active learning* aims to reduce this data requirement by studying the design of algorithms that can learn and generalize from a small carefully chosen subset of the training data [13, 40].

The two common formulations of active learning are *pool based* active learning, and *sequential (or streaming)* active learning. In the pool based setting [30], the learning algorithm has access to a large unlabeled set of data points, and the algorithm can ask for a subset of the data to be labeled. In contrast, in the sequential setting, data points arrive in a streaming manner, either adversarially or drawn i.i.d. from a distribution, and the algorithm must decide whether to query the label of a given point or not [14].

From a theoretical perspective, active learning has typically been studied under models inspired by the probably approximately correct (PAC) model of learning [41]. Here one assumes that there is a pre-specified class $\mathcal{H}$ of functions such that the *target* function mapping examples to their labels

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

either lies in $\mathcal{H}$ or has a good approximation inside the class. Given access to unlabeled samples generated i.i.d. from the distribution, the goal is to query for a small number of labels and produce a hypothesis of low error.

In the *parametric* setting, namely, when the class of functions $\mathcal{H}$ has finite VC-dimension (or finite disagreement coefficient) [21], the rate of convergence of active learning, i.e., the rate of decay of the regret as a function of the number of label queries (N), is of the form $\nu\, N^{-1/2} + e^{-\sqrt{N}}$, where ν is the population loss of the best function in class $\mathcal{H}$. This simple finding shows that active learning behaves like passive learning when $\nu > 0$, while very fast rates can only be achieved under low noise ($\nu \approx 0$) conditions. This has been worked out in, e.g., [19, 15, 5, 4, 6, 38].

While the parametric setting comes with methodological advantages, the above shows that in order to unleash the true power of active learning, two properties are desirable: (1) A better interplay between the input distribution and the label noise and, (2) a departure from the parametric setting leading us to consider wider classes of functions (so as to reduce the population loss ν to close to 0). To address the above, there has also been considerable theoretical work in recent years on non-parametric active learning [10, 33, 31]. However, these approaches suffer from the curse of dimensionality and do not lead to computationally efficient algorithms. A popular approach that has been explored empirically in recent works is to use Deep Neural Networks (DNNs) to perform active learning (e.g., [37, 26, 39, 3, 44]). While these works empirically demonstrate the power of the DNN-based approach to active learning, they do not come with provable guarantees. The above discussion raises the following question: *Is provable and computationally efficient active learning possible in non-parametric settings?*

We answer the above question in the affirmative by providing the first, to the best of our knowledge, computationally efficient algorithm for active learning based on Deep Neural Networks. Similar to non-parametric active learning, we avoid fixing a function class a-priori. However, in order to achieve computational efficiency, we instead propose to use over-parameterized DNNs, where the amount of over-parameterization depends on the input data at hand. We work in the sequential setting, and propose a simple active learning algorithm that forms an uncertainty estimate for the current data point based on the output of a DNN, followed by a gradient descent step to update the network parameters if the data point is queried. We show that under standard low-noise assumptions [32] our proposed algorithm achieves fast rates of convergence.

In order to analyze our algorithm, we use tools from the theory of Neural Tangent Kernel (NTK) approximation [24, 2, 18] that allows us to analyze the dynamics of gradient descent by considering a linearization of the network around random initialization. Since we study the non-parametric regime, the convergence rates of our algorithm depend on a data-dependent complexity term that is expected to be small in practical settings, but could be very large in worst-case scenarios. Furthermore, the algorithm itself needs an estimate of complexity term in order to form accurate uncertainty estimates. We show that one can automatically adapt to the magnitude of the unknown complexity term by designing a novel model selection algorithm inspired by recent works in model selection in multi-armed bandit settings [36, 35]. Yet, several new insights are needed to ensure that the model selection algorithm can simultaneously achieve low generalization error without spending a significant amount of budget on label queries.

2 Preliminaries and Notation

Let $\mathcal{X}$ denote the input space, $\mathcal{Y}$ the output space, and $\mathcal{D}$ an unknown distribution over $\mathcal{X} \times \mathcal{Y}$. We denote the corresponding random variables by x and y. We also denote by $\mathcal{D}_{\mathcal{X}}$ the marginal distribution of $\mathcal{D}$ over $\mathcal{X}$, and by $\mathcal{D}_{\mathcal{Y}|x_0}$ the conditional distribution of random variable y given $x = x_0$. Moreover, given a function f (sometimes called a hypothesis or a model) mapping $\mathcal{X}$ to $\mathcal{Y}$, the conditional *population loss* (often referred to as conditional *risk*) of f is denoted by $L(f \,|\, x)$, and defined as $L(f \,|\, x) = \mathbb{E}_{y \sim \mathcal{D}_{\mathcal{Y}|x}}[\ell(f(x), y) \,|\, x]$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is a *loss* function. For ease of presentation, we restrict to a binary classification setting with 0-1 loss, whence $\mathcal{Y} = \{-1, +1\}$, and $\ell(a, y) = \mathbb{1}\{a \neq y\} \in \{0, 1\}$, $\mathbb{1}\{\cdot\}$ being the indicator function of the predicate at argument. When clear from the surrounding context, we will omit subscripts like "$y \sim \mathcal{D}_{\mathcal{Y}|x}$" from probabilities and expectations.

We investigate a *non-parametric* setting of active learning where the conditional distribution of y given x is defined through an unknown function $h : \mathcal{X}^2 \to [0,1]$ such that

$$\mathbb{P}(y = 1 \,|\, x) = h((x,0)) \qquad \mathbb{P}(y = -1 \,|\, x) = h((0,x)) \;, \tag{1}$$

where $0 \in \mathcal{X}$, (x_1, x_2) denotes the concatenation (or pairing) of the two instances x_1 and x_2 (so that $(x,0)$ and $(0,x)$ are in $\mathcal{X}^2$) and, for all $x \in \mathcal{X}$ we have $h((x,0)) + h((0,x)) = 1$. We make no explicit assumptions on h, other than its well-behavedness w.r.t. the data $\{x_t\}_{t=1}^T$ at hand through the formalism of Neural Tangent Kernels (NTK) – see below. As a simple example, in the linear case, $\mathcal{X}$ is the d-dimensional unit ball, $h(\cdot,\cdot)$ is parametrized by an unknown unit vector $\theta \in \mathbb{R}^d$, and $h((x_1,x_2)) = \frac{1+\langle(\theta,-\theta),(x_1,x_2)\rangle}{2}$, so that $h((x,0)) = \frac{1+\langle\theta,x\rangle}{2}$ and $h((0,x)) = \frac{1-\langle\theta,x\rangle}{2}$, where $\langle\cdot,\cdot\rangle$ is the usual dot product in $\mathbb{R}^d$.

We consider a streaming setting of active learning where, at each round $t \in [T] = \{1,\dots,T\}$, a pair $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ is drawn i.i.d. from $\mathcal{D}$. The learning algorithm receives as input only x_t, and is compelled to both issue a prediction a_t for y_t and, at the same time, decide on-the-fly whether or not to observe y_t. These decisions can only be based on past observations. Let $\mathbb{E}_t$ denote the conditional expectation $\mathbb{E}[\cdot\,|(x_1,y_1)\dots,(x_{t-1},y_{t-1}),x_t]$, and we introduce the shorthand

$$x_{t,a} = \begin{cases} (x_t,0) & \text{if } a = 1 \\ (0,x_t) & \text{if } a = -1 \;. \end{cases}$$

Notice that with this notation $\mathbb{E}[\ell(a,y_t)\,|\,x_t] = 1 - h(x_{t,a})$, for all $a \in \mathcal{Y}$. We quantify the accuracy of the learner's predictions through its (pseudo) *regret*, defined as

$$R_T \;=\; \sum_{t=1}^T \Big(\mathbb{E}_t[\ell(a_t,y_t)\,|\,x_t] - \mathbb{E}[\ell(a_t^*,y_t)\,|\,x_t]\Big) \;=\; \sum_{t=1}^T \big(h(x_{t,a_t^*}) - h(x_{t,a_t})\big) \;,$$

where a_t^* is the Bayesian-optimal classifier on instance x_t, that is, $a_t^* = \arg\max_{a\in\mathcal{Y}} h(x_{t,a})$. Additionally, we are interested in bounding the number of labels N_T the algorithm decides to request. Our goal is to *simultaneously* bound R_T and N_T with high probability over the generation of the sample $\{(x_t,y_t)\}_{t=1,\dots,T}$.

Throughout this work, we consider the following common low-noise condition on the marginal distribution $\mathcal{D}_\mathcal{X}$ (Mammen-Tsybakov low noise condition [32]): There exist absolute constants $c > 0$, and $\alpha \geq 0$ such that for all $\epsilon \in (0,1/2)$ we have

$$\mathbb{P}\big(|h((x,0)) - \frac{1}{2}| < \epsilon\big) \leq c\,\epsilon^\alpha \;.$$

In particular, $\alpha = \infty$ gives the so-called *hard margin* condition $\mathbb{P}\big(|h((x,0)) - \frac{1}{2}| < \epsilon\big) = 0$. while, at the opposite extreme, exponent $\alpha = 0$ (and $c = 1$) results in *no assumptions whatsoever* on $\mathcal{D}_\mathcal{X}$. For simplicity, we shall assume throughout that the above low-noise condition holds for[1] $c = 1$.

Our techniques are inspired by the recent work [45] from which we also borrow some notation. We are learning the class of functions $\{h\}$ by means of fully connected neural networks

$$f(x,\theta) = \sqrt{m} W_n \sigma(\dots\sigma(W_1 x)) \;,$$

where σ is a ReLU activation function $\sigma(x) = \max\{0,x\}$, m is the width of the network and $n \geq 2$ is its depth. In the above, $\theta \in \mathbb{R}^p$ collectively denotes the set of weights $\{W_1, W_2, \dots, W_n\}$ of the network, where $p = m + 2md + m^2(n-2)$ is their number, and the input x at training time should be thought of as some $x_{t,a} \in \mathcal{X}^2$.

With any depth-n network and data points $\{x_{t,a}\}_{t=1,\dots,T,\,a=\pm1}$ we associate a depth-n NTK matrix as follows [24]. First, rename $\{x_{t,a}\}_{t=1,\dots,T,\,a=\pm1}$ as $\{x^{(i)}\}_{i=1,\dots,2T}$. Then define matrices

$$\widetilde{H}^{(1)} = \Big[H_{i,j}^{(1)}\Big]_{i,j=1}^{2T\times 2T} \qquad \Sigma^{(1)} = \Big[\Sigma_{i,j}^{(1)}\Big]_{i,j=1}^{2T\times 2T} \qquad \text{with} \qquad H_{i,j}^{(1)} = \Sigma_{i,j}^{(1)} = \langle x^{(i)}, x^{(j)}\rangle \;,$$

and then, for any $k \leq n$ and $i,j = 1,\dots,2T$, introduce the bivariate covariance matrix

$$A_{i,j}^{(k)} = \begin{bmatrix} \Sigma_{i,i}^{(k)} & \Sigma_{i,j}^{(k)} \\ \Sigma_{i,j}^{(k)} & \Sigma_{j,j}^{(k)} \end{bmatrix}$$

[1] A more general formulation requires the above to hold only for $\epsilon \leq \epsilon_0$, where $\epsilon_0 \in (0,1/2)$ is a third parameter. We shall omit this extra parameter from our presentation.

by which we recursively define

$$\Sigma_{i,j}^{(k+1)} = 2\mathbb{E}_{(u,v)\sim N(0,A_{i,j}^{(k)})}[\sigma(u)\sigma(v)]$$

and

$$\widetilde{H}_{i,j}^{(k+1)} = 2\widetilde{H}_{i,j}^{(k)}\mathbb{E}_{(u,v)\sim N(0,A_{i,j}^{(k)})}[\mathbb{1}\{u \geq 0\}\,\mathbb{1}\{v \geq 0\}] + \Sigma_{i,j}^{(k+1)} \,.$$

The $2T \times 2T$-dimensional matrix $H = \frac{1}{2}(\widetilde{H}^{(n)} + \Sigma^{(n)})$ is called the Neural Tangent Kernel (NTK) matrix of depth n (and infinite width) over the set of points $\{x_{t,a}\}_{t=1,\dots,T,\, a=\pm 1}$. The reader is referred to [24] for more details on NTK.

In order to avoid heavy notation, we assume $||x_t|| = 1$ for all t. Matrix H is positive semi-definite by construction but, as is customary in the NTK literature (e.g., [2, 9, 17]), we assume it is actually positive definite (hence invertible) with smallest eigenvalue $\lambda_0 > 0$. This is a mild assumption that can be shown to hold if no two vectors x_t are aligned to each other.

We measure the complexity of the function h at hand in a way similar to [45]. Using the same rearrangement of $\{x_{t,a}\}_{t=1,\dots,T,\, a=\pm 1}$ into $\{x^{(i)}\}_{i=1,\dots,2T}$ as above, let $\mathbf{h}$ be the $2T$-dimensional (column) vector whose i-th component is $h(x^{(i)})$. Then, we define the complexity $S_{T,n}(h)$ of h over $\{x_{t,a}\}_{t=1,\dots,T,\, a=\pm 1}$ w.r.t. an NTK of depth n as $S_{T,n}(h) = \sqrt{\mathbf{h}^\top H^{-1}\mathbf{h}}$. Notice that this notion of (data-dependent) complexity is consistent with the theoretical findings of [2], who showed that for a two-layer network the bound on the generalization performance is dominated by $\mathbf{y}^\top H^{-1}\mathbf{y}$, where $\mathbf{y}$ is the vector of labels. Hence if $\mathbf{y}$ is aligned with the top eigenvectors of H the learning problem becomes easier. In our case, vector $\mathbf{h}$ plays the role of vector $\mathbf{y}$. Also observe that $S_{T,n}^2(h)$ can in general be as big as linear in T (in which case learning becomes hopeless with our machinery). In the special case where h belongs to the RKHS induced by the NTK, one can upper bound $S_{T,n}(h)$ by the norm of h in the RKHS.

The complexity term $S_{T,n}(h)$ is typically *unknown* to the learning algorithm, and it plays a central role in both regret and label complexity guarantees. Hence the algorithm needs to *learn* this value as well during its online functioning. Apparently, this aspect of the problem has been completely overlooked by [45] (as well as by earlier references on contextual bandits in RKHS, like [12]), where a (tight) upper bound on $S_{T,n}(h)$ is assumed to be available in advance. We will cast the above as a *model selection* problem in active learning, where we adapt and largely generalize to active learning the regret balancing technique from [36, 35].

In what follows, we use the short-hand $g(x;\theta) = \nabla_\theta f(x,\theta)$ and, for a vector $g \in \mathbb{R}^p$ and matrix $Z \in \mathbb{R}^{p\times p}$, we often write $\sqrt{g^\top Z g}$ as $||g||_Z$, so that $S_{T,n}(h) = ||\mathbf{h}||_{H^{-1}}$.

2.1 Related work

The main effort in theoretical works in active learning is to obtain rates of convergence of the population loss of the hypothesis returned by the algorithm as a function of the number N of requested labels. We emphasize that most of these works, that heavily rely on approximation theory, are *not* readily comparable to ours, since our goal here is not to approximate h through a DNN on the entire input domain, but only on the data at hand.

As we recalled in the introduction, in the *parametric* setting the convergence rates of the regret are of the form $\nu\, N^{-1/2} + e^{-\sqrt{N}}$, where ν is the population loss of the best function in class $\mathcal{H}$. Hence, active learning rates behave like the passive learning rate $N^{-1/2}$ when $\nu > 0$, while fast rates can only be achieved under very low noise ($\nu \approx 0$) conditions. In this respect, relevant references include [20, 27] where, e.g., in the realizable case (i.e., when the Bayes optimal classifier lies in $\mathcal{H}$), minimax active learning rates of the form $N^{-\frac{\alpha+1}{2}}$ are shown to hold for adaptive algorithms that do not know beforehand the noise exponent α. In non-parametric settings, a comprehensive set of results has been obtained by [31], which builds on and significantly improves over earlier results from [33]. Both papers work under smoothness (Holder continuity/smoothness) assumptions. In addition, [33] requires $\mathcal{D}_\mathcal{X}$ to be (quasi-)uniform on $\mathcal{X} = [0,1]^d$. In [31] the minimax active learning rate $N^{-\frac{\beta(\alpha+1)}{2\beta+d}}$ is shown to hold for β-Holder classes, where exponent β plays the role of the complexity of the class of functions to learn, and d is the input dimension. This algorithm is adaptive to the complexity parameter β, and is therefore performing a kind of model selection. Notice that minimax rates in the

parametric regime are recovered by setting $\beta \to \infty$. Of a somewhat similar flavor is an earlier result by [27], where a convergence rate of the form $N^{-\frac{\alpha+1}{2+\kappa\alpha}}$ is shown, being κ the metric entropy of the class (again, a notion of complexity). A refinement of the results in [31] has recently been obtained by [34] where, following [11], a more refined notion of smoothness for the Bayes classifier is adopted which, however, also implies more restrictive assumptions on the marginal distribution $\mathcal{D}_{\mathcal{X}}$.

As opposed to those bounds, our bounds are *data-dependent*, in that all relevant quantities appearing in the bounds will be random variables depending on the data at hand (which are themselves random). One may attempt to turn these into *data-independent* results (like in most of the papers we cited above) by, e.g., establishing bounds on that hold in expectation or with high probability over the random draw of the data, but this theory is currently unavailable in the NTK literature (as far as we know). Very recently some results have appeared for certain special cases, see [23] for example. But such results are too embryonic in nature to allow us a full-fledged comparison.

Model selection of the scale of a Nearest-Neighbor-based active learning algorithm is also performed in [28], whose main goal is to achieve data-dependent rates based on the noisy-margin properties of the random sample at hand, rather than those of the marginal distribution. Their active learning rates are not directly comparable to ours and, unlike our paper, the authors work in a *pool-based* scenario, where all unlabeled points are available beforehand. Finally, an interesting investigation in active learning for over-parametrized and interpolating regimes is contained in [25]. The paper collects a number of interesting insights in active learning for 2-layer Neural Networks and Kernel methods, but it restricts to either uniform distributions on the input space or cases of well-clustered data points, with no specific regret and query complexity guarantees, apart from very special (though insightful) cases.

3 Basic Algorithm

Our first algorithm (Algorithm 1) uses randomly initialized, but otherwise frozen, network weights (a more refined algorithm where the network weights are updated incrementally is described and analyzed in the appendix). Algorithm 1 is an adaptation to active learning of the neural contextual bandit algorithm of [45], and shares similarities with an earlier selective sampling algorithm analyzed in [16] for the linear case. The algorithm generates network weights θ_0 by independently sampling from Gaussian distributions of appropriate variance, and then uses θ_0 to stick with a gradient mapping $\phi(\cdot)$ which will be kept frozen from beginning to end. The algorithm also takes as input the complexity parameter $S = S_{T,n}(h)$ of the underlying function h satisfying (1). We shall later on remove the assumption of the prior knowledge of $S_{T,n}(h)$. In particular, removing the latter, turns out to be quite challenging from a technical standpoint, and gives rise to a complex online model selection algorithms for active learning in non-parametric regimes.

At each round t, Algorithm 1 receives an instance $x_t \in \mathcal{X}$, and constructs the two augmented vectors $x_{t,1} = (x_t, 0)$ and $x_{t,-1} = (0, x_t)$ (intuitively corresponding to the two "actions" of a contextual bandit algorithm). The algorithm predicts the label y_t associated with x_t by maximizing over $a \in \mathcal{Y}$ an upper confidence index $U_{t,a}$ stemming from the linear approximation $h(x_{t,a}) \approx \sqrt{m}\langle \phi(x_{t,a}), \theta_{t-1} - \theta_0 \rangle$ subject to ellipsoidal constraints $\mathcal{C}_{t-1}$, as in standard contextual bandit algorithms operating with the frozen mapping $\phi(\cdot)$. In addition, in order to decide whether or not to query label y_t, the algorithm estimates its own uncertainty by checking to what extent U_{t,a_t} is close to $1/2$. This uncertainty level is ruled by the time-varying threshold B_t, which is expected to shrink to 0 as time progresses. Notice that B_t is a function of γ_{t-1}, which in turn includes in its definition the complexity parameter S. Finally, if y_t is revealed, the algorithm updates its least-squares estimator θ_t by a rank-one adjustment of matrix Z_t and an additive update to the bias vector b_t. No update is taking place if the label is not queried. The following is our initial building block.[2]

Theorem 1. *Let Algorithm 1 be run with parameters δ, S, m, and n on an i.i.d. sample $(x_1, y_1), \ldots, (x_T, y_T) \sim \mathcal{D}$, where the marginal distribution $\mathcal{D}_{\mathcal{X}}$ fulfills the low-noise condition with exponent $\alpha \geq 0$ w.r.t. a function h that satisfies (1) and such that $\sqrt{2}S_{T,n}(h) \leq S$. If $m = poly(T, n, \lambda_0^{-1}, \log(1/\delta))$, then with probability at least $1 - \delta$ the cumulative regret R_T and*

[2] All proofs are in the appendix.

Algorithm 1: Frozen NTK Selective Sampler.

Input: Confidence level δ, complexity parameter S, network width m, and depth n .

Initialization:

- Generate each entry of W_k independently from $\mathcal{N}(0, 2/m)$, for $k \in [n-1]$, and each entry of W_n independently from $\mathcal{N}(0, 1/m)$;
- Define $\phi(x) = g(x; \theta_0)/\sqrt{m}$, where $\theta_0 = \langle W_1, \dots, W_n \rangle \in \mathbb{R}^p$ is the (frozen) weight vector of the neural network so generated;
- Set $Z_0 = I \in \mathbb{R}^{p \times p},\ b_0 = 0 \in \mathbb{R}^p$.

for $t = 1, 2, \dots, T$

Observe instance $x_t \in \mathcal{X}$ and build $x_{t,a} \in \mathcal{X}^2$, for $a \in \mathcal{Y} = \{-1, +1\}$

Set $\mathcal{C}_{t-1} = \{\theta : \|\theta - \theta_{t-1}\|_{Z_{t-1}} \leq \frac{\gamma_{t-1}}{\sqrt{m}}\}$, with $\gamma_{t-1} = \sqrt{\log\det Z_{t-1} + 2\log(1/\delta)} + S$

Set

$$U_{t,a} = \sqrt{m} \max_{\theta \in \mathcal{C}_{t-1}} \langle \phi(x_{t,a}), \theta - \theta_0 \rangle = \sqrt{m} \langle \phi(x_{t,a}), \theta_{t-1} - \theta_0 \rangle + \gamma_{t-1} \|\phi(x_{t,a})\|_{Z_{t-1}^{-1}}$$

Predict $a_t = \arg\max_{a \in \mathcal{Y}} U_{t,a}$

Set $I_t = \mathbb{1}\{|U_{t,a_t} - 1/2| \leq B_t\} \in \{0, 1\}$ with $B_t = B_t(S) = 2\gamma_{t-1} \|\phi(x_{t,a_t})\|_{Z_{t-1}^{-1}}$

if $I_t = 1$

Query $y_t \in \mathcal{Y}$, and set loss $\ell_t = \ell(a_t, y_t)$

Update

$$Z_t = Z_{t-1} + \phi(x_{t,a_t})\phi(x_{t,a_t})^\top$$
$$b_t = b_{t-1} + (1 - \ell_t)\phi(x_{t,a_t})$$
$$\theta_t = Z_t^{-1} b_t / \sqrt{m} + \theta_0$$

else $Z_t = Z_{t-1},\ b_t = b_{t-1},\ \theta_t = \theta_{t-1},\ \gamma_t = \gamma_{t-1},\ \mathcal{C}_t = \mathcal{C}_{t-1}$.

the total number of queries N_T are simultaneously upper bounded as follows:

$$R_T = O\left(L_H^{\frac{\alpha+1}{\alpha+2}} \left(L_H + \log(1/\delta) + S^2 \right)^{\frac{\alpha+1}{\alpha+2}} T^{\frac{1}{\alpha+2}} + \log(\log T/\delta) \right)$$
$$N_T = O\left(L_H^{\frac{\alpha}{\alpha+2}} \left(L_H + \log(1/\delta) + S^2 \right)^{\frac{\alpha}{\alpha+2}} T^{\frac{2}{\alpha+2}} + \log(\log T/\delta) \right),$$

where $L_H = \log\det(I + H)$, H being the NTK matrix of depth n over the set of points $\{x_{t,a}\}_{t=1,\dots,T,\, a=\pm 1}$.

The above bounds depend, beyond time horizon T, on three relevant quantities: the noise level α, the complexity parameters S and the log-determinant quantity L_H. Notice that, whereas S essentially quantifies the complexity of the function h to be learned, L_H measures instead the complexity of the NTK itself, hence somehow quantifying the complexity of the function space we rely upon in learning h. It is indeed instructive to see how the bounds in the above theorem vary as a function of these quantities. First, as expected, when $\alpha = 0$ we recover the usual regret guarantee $R_T = O(\sqrt{T})$, more precisely a bound of the form $R_T = O((L_H + \sqrt{L_H}S)\sqrt{T})$, with the trivial label complexity $N_T = O(T)$. At the other extreme, when $\alpha \to \infty$ we obtain the guarantees $R_T = N_T = O(L_H(L_H + S^2))$. In either case, if h is "too complex" when projected onto the data, that is, if $S_{T,n}^2(h) = \Omega(T)$, then all bounds become vacuous.[3] At the opposite end of the spectrum, if $\{h\}$ is simple, like a class of linear functions with bounded norm in a d-dimensional space, and the network depth n is 2 then $S_{T,n}(h) = O(1)$, and $L_H = O(d \log T)$, hence recovering the rates reported in [16] for the linear case. The quantity L_H is tightly related to the decaying rate of the eigenvalues of the NTK matrix H, and is poly-logarithmic in T in several important cases [42]. One relevant example is discussed in [43], which relies on the spectral characterization of NTK in [7, 8]: If $n = 2$ and all points $x^{(i)}$ concentrate on a d_0-dimensional subspace of the RKHS spanned by the NTK, then $L_H = O(d_0 \log T)$.

[3] The same happens, e.g., to the regret bounds in [45].

It is also important to stress that, via a standard online-to-batch conversion, the result in Theorem 1 can be turned to a compelling guarantee in a traditional statistical learning setting, where the goal is to come up at the end of the T rounds with a hypothesis f whose population loss $L(f) = \mathbb{E}_{x\sim D_{\mathcal{X}}}[L(f \,|\, x)]$ exceeds the Bayes optimal population loss $\mathbb{E}_{x_t\sim D_{\mathcal{X}}}[h(x_{t,a_t^*})] = \mathbb{E}_{x_t\sim D_{\mathcal{X}}}[\max\{h(x_{t,1}), h(x_{t,-1})\}]$ by a vanishing quantity. Following [16], this online-to-batch algorithm will simply run Algorithm 1 by sweeping over the sequence $\{(x_t, y_t)\}_{t=1,\ldots,T}$ only once, and pick one function uniformly at random among the sequence of predictors generated by Algorithm 1 during its online functioning, that is, among the sequence $\{U_t(x)\}_{t=1,\ldots,T}$, where $U_t(x) = \arg\max_{a\in\mathcal{Y}} \max_{\theta\in\mathcal{C}_{t-1}} \langle \phi(x_{\cdot,a}), \theta - \theta_0\rangle$, with $x_{\cdot,1} = (x, 0)$ and $x_{\cdot,-1} = (0, x)$. This randomized algorithm enjoys the following high-probability excess risk guarantee:[4]

$$\mathbb{E}_{t\sim\text{unif}(T)}[L(U_t)] - \mathbb{E}_{x_t\sim D_{\mathcal{X}}}[h(x_{t,a_t^*})] = O\left(\left(\frac{L_H\Big(L_H + \log(1/\delta) + S^2\Big)}{T}\right)^{\frac{\alpha+1}{\alpha+2}} + \frac{\log\log(T/\delta)}{T}\right).$$

Combining with the guarantee on the number of labels N_T from Theorem 1 (and disregarding log factors), this allows us to conclude that the above excess risk can be bounded as a function of N_T as

$$\Big(\frac{L_H(L_H + S^2)}{N_T}\Big)^{\frac{\alpha+1}{2}}, \tag{2}$$

where $L_H(L_H + S^2)$ plays the role of a (compound) complexity term projected onto the data $x_1, \ldots, x_T$ at hand. When restricting to VC-classes, the convergence rate $N_T^{-\frac{\alpha+1}{2}}$ is indeed the best rate (*minimax* rate) one can achieve under the Mammen-Tsybakov low-noise condition with exponent α (see, e.g., [10, 20, 27, 16]).

Yet, since we are not restricting to the parametric case, both L_H and, more importantly, S^2 can be a function of T. In such cases, the generalization bound in (2) can still be expressed as a function of N_T alone, For instance, when L_H is poly-logarithmic in T and $S^2 = O(T^\beta)$, for some $\beta \in [0, 1)$, one can easily verify that (2) takes the form $N_T^{-\frac{(1-\beta)(\alpha+1)}{2+\beta\alpha}}$ (again, up to log factors).

In Section A.3 of the appendix, we extend all our results to the case where the network weights are not frozen, but are updated on the fly according to a gradient descent procedure. In this case, in Algorithm 1 the gradient vector $\phi(x) = g(x; \theta_0)/\sqrt{m}$ will be replaced by $\phi_t(x) = g(x; \theta_{t-1})/\sqrt{m}$, where θ_t is not the linear-least squares estimator $\theta_t = Z_t^{-1} b_t/\sqrt{m} + \theta_0$, as in Algorithm 1, but the result of the DNN training on the labeled data $\{(x_k, y_k) : k \le t,\, I_k = 1\}$ gathered so far.

4 Model Selection

Our model selection algorithm is described in Algorithm 2. The algorithm operates on a pool of *base learners* of Frozen NTK selective samplers like those in Algorithm 1, each member in the pool being parametrized by a pair of parameters (S_i, d_i), where S_i plays the role of the (unknown) complexity parameter $S_{T,n}(h)$ (which was replaced by S in Algorithm 1), and d_i plays the role of an (a-priori unknown) upper bound on the relevant quantity $\sum_{t\in T\,:\, i_t=i} \frac{1}{2} \wedge I_{t,i} B_{t,i}^2$ that is involved in the analysis (see Lemma 5 and Lemma 7 in Appendix A.1). This quantity will at the end be upper bounded by a term of the form $L_H(L_H + \log(T/\delta) + S_{T,n}^2(h))$ whose components L_H and $S_{T,n}^2(h)$ are initially unknown to the algorithm.

Algorithm 2 maintains over time a set $\mathcal{M}_t$ of active base learners, and a probability distribution $\boldsymbol{p}_t$ over them. This distribution remains constant throughout a sequence of rounds between one change to $\mathcal{M}_t$ and the next. We call such sequence of rounds an *epoch*. Upon observing x_t, Algorithm 2 selects which base learner to rely upon in issuing its prediction a_t and querying the label y_t, by drawing base learner $i_t \in \mathcal{M}_t$ according to $\boldsymbol{p}_t$.

Then Algorithm 2 undergoes a series of carefully designed elimination tests which are meant to rule out mis-specified base learners, that is, those whose associated parameter S_i is likely to be smaller than $S_{T,n}(h)$, while retaining those such that $S_i \ge S_{T,n}(h)$. These tests will help keep both the regret bound and the label complexity of Algorithm 2 under control. Whenever, at the end of some

[4] Observe that this is a *data-dependent* bound, in that the RHS is random variable. This is because both L_H and S may depend on $x_1, \ldots, x_T$.

round t, any such test triggers, that is, when it happens that $|\mathcal{M}_{t+1}| < |\mathcal{M}_t|$ at the end of the round, a new epoch begins, and the algorithm starts over with a fresh distribution $\boldsymbol{p}_{t+1} \neq \boldsymbol{p}_t$.

The first test ("disagreement test") restricts to all active base learners that would not have requested the label if asked. As our analysis for the base selective sampler (see Lemma 8 in Appendix A.1) shows that a well-specified base learner does not suffer (with high probability) any regret on non-queried rounds, any disagreement among them reveals mis-specification, thus we eliminate in pairwise comparison the base learner that holds the smaller S_i parameter. The second test ("observed regret test") considers the regret behavior of each pair of base learners $i, j \in \mathcal{M}_t$ on the rounds $k \leq t$ on which i was selected ($i_k = i$) and requested the label ($I_{k,i} = 1$), but j would not have requested if asked ($I_{k,j} = 0$), and the predictions of the two happened to disagree on that round ($a_{k,i} \neq a_{k,j}$). The goal here is to eliminate base learners whose cumulative regret is likely to exceed the regret of the smallest well-specified learner, while ensuring (with high probability) that any well-specified base learner i is not removed from the pool. In a similar fashion, the third test ("label complexity test") is aimed at keeping under control the label complexity of the base learners in the active pool $\mathcal{M}_t$. Finally, the last test ("d_i test") simply checks whether or not the candidate value d_i associated with base learner i remains a valid (and tight) upper bound on $L_H(L_H + S^2_{T,n}(h))$.

Notice that the sampling distribution $\boldsymbol{p}_t$ plays base learners with small d_i more often than learners with large d_i. Note also that d_i is exactly the (instance-dependent) factor in the cumulative regret and label complexity bounds for base learners that are well-specified. This means that base learners with lower regret are chosen more frequently than base learners that accumulate regret quicker (and similarly for label complexity). In fact, the sampling distribution is chosen so that the total contribution to the cumulative regret of each base learner is roughly equal. As a consequence, the total cumulative regret of Algorithm 2 is at most M (number of base learners) times the regret of each base learner, and the best base learner in particular, which is a key property for achieving the guarantees in Theorem 2 below. Of course, this only works when the base learners are well-specified but the four tests in Algorithm 2 ensure that all other learners are eventually eliminated.

We have the following result, whose proof is contained in Appendix A.2.

Theorem 2. *Let Algorithm 2 be run with parameters δ, $\gamma \leq \alpha$ with a pool of base learners $\mathcal{M}_1$ of size M on an i.i.d. sample $(x_1, y_1), \dots, (x_T, y_T) \sim \mathcal{D}$, where the marginal distribution $\mathcal{D}_\mathcal{X}$ fulfills the low-noise condition with exponent $\alpha \geq 0$ w.r.t. a function h that satisfies (1) and having complexity $S_{T,n}(h)$. Let also $\mathcal{M}_1$ contain at least one base learner i such that $\sqrt{2}S_{T,n}(h) \leq S_i \leq 2\sqrt{2}S_{T,n}(h)$ and $d_i = \Theta(L_H(L_H + \log(M/\delta) + S^2_{T,n}(h)))$, where $L_H = \log\det(I + H)$, being H the NTK matrix of depth n over the set of points $\{x_{t,a}\}_{t=1,\dots,T,\, a=\pm 1}$. If $m = poly(T, n, \lambda_0^{-1}, \log(1/\delta))$, then with probability at least $1 - \delta$ the cumulative regret R_T and the total number of queries N_T are simultaneously upper bounded as follows:*

$$R_T = O\left(M\left(L_H\big(L_H + \log(M/\delta) + S^2_{T,n}(h)\big)\right)^{\gamma+1} T^{\frac{1}{\gamma+2}} + M\,L(T,\delta)\right)$$

$$N_T = O\left(M\left(L_H\big(L_H + \log(M/\delta) + S^2_{T,n}(h)\big)\right)^{\frac{\gamma}{\gamma+2}} T^{\frac{2}{\gamma+2}} + M\,L(T,\delta)\right),$$

where $L(T,\delta)$ is the logarithmic term defined at the beginning of Algorithm 2's pseudocode.

We run Algorithm 2 with the pool $\mathcal{M}_1 = \{(S_{i_1}, d_{i_2})\}$, where $S_{i_1} = 2^{i_1}$, $i_1 = 0, 1, \dots, O(\log T)$ and $d_{i_2} = 2^{i_2}$, $i_2 = 0, 1, \dots, O(\log T)$, ensuring[5] the existence of a pair (i_1, i_2) such that

$$\sqrt{2}S_{T,n}(h) \leq S_{i_1} \leq 2\sqrt{2}S_{T,n}(h)$$

and

$$L_H\big(L_H + \log(M/\delta) + S^2_{T,n}(h)\big) \leq d_{i_2} \leq 2L_H\big(L_H + \log(M/\delta) + S^2_{T,n}(h)\big)\,.$$

Hence the resulting error due to the discretization is just a constant factor, while the resulting number M of base learners is $O(\log^2 T)$.

Theorem 2 allows us to conclude that running Algorithm 2 on the above pool of copies of Algorithm 1 yields guarantees that are similar to those obtained by running a single instance of Algorithm 1 with

[5]Notice that the bounds in Theorem 2 become vacuous if either $S_{T,n}(h)$ or L_H are $\Theta(\sqrt{T})$, hence we are only interested in making indices i_1 and i_2 reach a value which is at most logarithmic in T.

Algorithm 2: Frozen NTK Selective Sampler with Model Selection.

Input: Confidence level δ; probability parameter $\gamma \geq 0$; pool of base learners $\mathcal{M}_1$, each identified with a pair (S_i, d_i); number of rounds T.

Set $L(t,\delta) = \log \frac{5.2\log(2t)^{1.4}}{\delta}$

for $t = 1, 2, \ldots, T$

Observe instance $x_t \in \mathcal{X}$ and build $x_{t,a} \in \mathcal{X}^2$, for $a \in \mathcal{Y} = \{-1, +1\}$

for $i \in \mathcal{M}_t$

Set $I_{t,i} \in \{0,1\}$ as the indicator of whether base learner i *would* ask for label on x_t

Set $a_{t,i} \in \mathcal{Y}$ as the prediction of base learner i on x_t

Let $B_{t,i} = B_{t,i}(S_i)$ denote the query threshold of base learner i (from Algorithm 1)

Select base learner $i_t \sim \boldsymbol{p}_t = (p_{t,1}, p_{t,2}, \ldots, p_{t,|\mathcal{M}_t|})$, where

$$p_{t,i} = \begin{cases} \frac{d_i^{-(\gamma+1)}}{\sum_{j\in\mathcal{M}_t} d_j^{-(\gamma+1)}}, & \text{if } i \in \mathcal{M}_t \\ 0, & \text{otherwise} \end{cases}$$

Predict $a_t = a_{t,i_t}$

if $I_{t,i_t} = 1$

Query label $y_t \in \mathcal{Y}$ and send (x_t, y_t) to base learner i_t

$\mathcal{M}_{t+1} = \mathcal{M}_t$

Set $\mathcal{N}_t = \{i \in \mathcal{M}_t \colon I_{t,i} = 0\}$ `// (1) Disagreement test`

for *all pairs of base learners* $i, j \in \mathcal{N}_t$ *that disagree in their prediction* ($a_{t,i} \neq a_{t,j}$)

Eliminate all learners with smaller S: $\mathcal{M}_{t+1} = \{m \in \mathcal{M}_{t+1} \colon S_m > \min\{S_i, S_j\}\}$

for *all pairs of base learners* $i, j \in \mathcal{M}_t$ `// (2) Observed regret test`

Consider rounds where the chosen learner i requested the label but j did not, and i and j disagree in their prediction:

$$\mathcal{V}_{t,i,j} = \{k \in [t] \colon i_k = i, I_{k,i} = 1, I_{k,j} = 0, a_{k,i} \neq a_{k,j}\}$$

if $\sum_{k\in\mathcal{V}_{t,i,j}} (\mathbb{1}\{a_{k,i} \neq y_k\} - \mathbb{1}\{a_{k,j} \neq y_k\}) > \sum_{k\in\mathcal{V}_{t,i,j}} (1 \wedge B_{k,i}) + 1.45\sqrt{|\mathcal{V}_{t,i,j}| L(|\mathcal{V}_{t,i,j}|, \delta)}$

Eliminate base learner i: $\mathcal{M}_{t+1} = \mathcal{M}_{t+1} \setminus \{i\}$

for $i \in \mathcal{M}_t$ `// (3) Label complexity test`

Consider rounds where base learner i was played: $\mathcal{T}_{t,i} = \{k \in [t] \colon i_k = i\}$

if

$$\sum_{k\in\mathcal{T}_{t,i}} I_{k,i} > \inf_{\epsilon\in(0,1/2]} \left(3\epsilon^{\gamma} |\mathcal{T}_{t,i}| + \frac{1}{\epsilon^2} \sum_{k\in\mathcal{T}_{t,i}} I_{k,i} B_{k,i}^2 \wedge \frac{1}{4} \right) + 2L(|\mathcal{T}_{t,i}|, \delta/(M \log_2(12t)))$$

Eliminate base learner i: $\mathcal{M}_{t+1} = \mathcal{M}_{t+1} \setminus \{i\}$

for $i \in \mathcal{M}_t$ `// (4)` d_i `test`

if $\sum_{k\in\mathcal{T}_{t,i}} (\frac{1}{2} \wedge I_{k,i} B_{k,i}^2) > 8 d_i$

Eliminate base learner i: $\mathcal{M}_{t+1} = \mathcal{M}_{t+1} \setminus \{i\}$

$S = \sqrt{2} S_{T,n}(h)$, that is, as if the complexity parameter $S_{T,n}(h)$ were known beforehand. Yet, this model selection guarantee comes at a price, since Algorithm 2 needs to receive as input the noise exponent α (through parameter $\gamma \leq \alpha$) in order to correctly shape its label complexity test.

The very same online-to-batch conversion mentioned in Section 3 can be applied to Algorithm 2. Again, combining with the bound on the number of labels and disregarding log factors, this gives us a high probability excess risk bound of the form

$$\left(\frac{\left[L_H \left(L_H + S_{T,n}^2(h) \right) \right]^{\frac{3\alpha+2}{\alpha+2}}}{N_T} \right)^{\frac{\alpha+1}{2}}, \tag{3}$$

provided $\gamma = \alpha$. Following the same example as at the end of Section 3, when L_H is poly-logarithmic in T and $S^2 = O(T^\beta)$, for some $\beta \in [0, 1)$, one can verify that (3) is of the form $N_T^{-\frac{(1-\beta(\alpha+1))(\alpha+1)}{2+\beta\alpha}}$ (up to log factors), which converges for $\beta < 1/(\alpha + 1)$. Hence, compared to (2) we can ensure convergence in a more restricted set of cases.

Section A.3 in the appendix contains the extension of our model selection procedure to the case where the network weights are themselves updated.

5 Conclusions and Work in Progress

We have presented a rigorous analysis of selective sampling and active learning in general non-parametric scenarios, where the complexity of the Bayes optimal predictor is evaluated on the data at hand as a fitting measure with respect to the NTK matrix of a given depth associated with the same data. This complexity measure plays a central role in the level of uncertainty the algorithm assigns to labels (the higher the complexity the higher the uncertainty, hence the more labels are queried). Yet, since this is typically an unknown parameter of the problem, special attention is devoted to designing and analyzing a model selection technique that adapts to this unknown parameter.

In doing so, we borrowed tools and techniques from Neural Bandits [45, 43], selective sampling (e.g., [16]), and online model selection in contextual bandits [36, 35], and combined them together in an original and non-trivial manner.

We proved regret and label complexity bounds that recover known minimax rates in the parametric case, and extended such results well beyond the parametric setting achieving favorable guarantees that cannot easily be compared to available results in the literature of active learning in non-parametric settings. One distinctive feature of our proposed technique is that it gives rise to efficient and manageable algorithms for modular DNN architecture design and deployment.

We conclude by mentioning a few directions we are currently exploring:

1. We are trying to get rid of the prior knowledge of α in the model selection Algorithm 2. This may call for a slightly more refined balancing technique that jointly involves $S_{T,n}(h)$ and α itself.
2. Regardless of whether α is available, it would be nice to improve the dependence on $\gamma = \alpha$ in the regret bound of Theorem 2. This would ensure convergence of the generalization bound as $N_T \to \infty$ when $S_{T,n}(h)^2 = T^\beta$, for all $\beta \in [0, 1)$. We conjecture that this is due to a suboptimal design of our balancing mechanism for model selection in Algorithm 2.
3. We are investigating links between the complexity measure $S_{T,n}(h)$ and the smoothness properties of the (Bayes) regression function h with respect to the NTK kernel (of a given depth n).

References

[1] Y. Abbasi-yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pages 2312–2320. Curran Associates, Inc., 2011.

[2] S. Arora, S. S. Du, W. Hu, Z. Li, R. Salakhutdinov, and R. Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[3] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[4] M. Balcan, A. Beygelzimer, and J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

[5] N. Balcan, S. Hanneke, and J. Wortman. The true sample complexity of active learning. In *COLT*, 2008.

[6] A. Beygelzimer, S. Dasgupta, and J. Langford. Importance weighted active learning. In *ICML*, 2009.

[7] A. Bietti and J. Mairal. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[8] Y. Cao, Z. Fang, Y. Wu, D. Zhou, and Q. Gu. Towards understanding the spectral bias of deep learning. In *arXiv:1912.01198*, 2019.

[9] Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[10] R. Castro and R. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

[11] K. Chaudhuri and S. Dasgupta. Rates of convergence for nearest neighbor classification. In *Advances in Neural Information Processing Systems*, pages 3437–3445, 2014.

[12] S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[13] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

[14] I. Dagan and S. P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995.

[15] S. Dasgupta, D. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems*, 2007.

[16] O. Dekel, C. Gentile, and K. Sridharan. Selective sampling and active learning from single and multiple teachers. *J. Mach. Learn. Res.*, 13(1), 2012.

[17] S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, page 1675–1685, 2019.

[18] S. Du, J. Lee, Y. Tian, A. Singh, and B. Poczos. Gradient descent learns one-hidden-layer CNN: Don't be afraid of spurious local minima. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1339–1348. PMLR, 2018.

[19] S. Hanneke. A bound on the label complexity of agnostic active learning. In *ICML*, 2007.

[20] S. Hanneke. Adaptive rates of convergence in active learning. In *Proc. of the 22th Annual Conference on Learning Theory*, 2009.

[21] S. Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

[22] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *arXiv preprint arXiv:1810.08240*, 2018.

[23] D. Hsu, C. Sanford, R. Servedio, and E. V. Vlatakis-Gkaragkounis. On the approximation power of two-layer networks of random relus. *arXiv preprint arXiv: 2102.02336*, 2021.

[24] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Advances in neural information processing systems*, page 8571–8580. MIT Press, 2018.

[25] M. Karzand and R. Nowak. Maximin active learning in overparameterized model classes. In *arXiv:1905.12782v2*. 2020.

[26] A. Kirsch, J. Van Amersfoort, and Y. Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *arXiv preprint arXiv:1906.08158*, 2019.

[27] V. Koltchinskii. Rademacher complexities and bounding the excess risk of active learning. *Journal of Machine Learning Research*, 11:2457–2485, 2010.

[28] A. Kontorovich, S. Sabato, and R. Urner. Active nearest-neighbor learning in metric spaces. In *Advances in Neural Information Processing Systems*, pages 856–864, 2016.

[29] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[30] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.

[31] C. A. Locatelli A. and S. Kpotufe. Adaptivity to noise parameters in nonparametric active learning. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1383–1416, 2017.

[32] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[33] S. Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13:67–90, 2012.

[34] B. Njike and X. Siebert. Nonparametric adaptive active learning under local smoothness condition. In *arxiv: 2102.11077*. 2021.

[35] A. Pacchiano, C. Dann, G. C., and P. Bartlett. Regret bound balancing and elimination for model selection in bandits and RL. *arXiv preprint arXiv:2012.13045*, 2020.

[36] A. Pacchiano, M. Phan, Y. Abbasi Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. In *Advances in Neural Information Processing Systems*, volume 33, pages 10328–10337. Curran Associates, Inc., 2020.

[37] R. Pop and P. Fulop. Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles. *arXiv preprint arXiv:1811.03897*, 2018.

[38] M. Raginsky and A. Rakhlin. Lower bounds for passive and active learning. In *Advances in Neural Information Processing Systems*, 2011.

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

[40] B. Settles. Active learning literature survey. 2009.

[41] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[42] M. Valko, N. Korda, R. Munos, I. Flaounas, and N. Cristianini. Finite-time analysis of kernelised contextual bandits. In *arxiv:1309.6869*. 2013.

[43] W. Zhang, D. Zhou, L. Li, and Q. Gu. Neural thompson sampling. In *arXiv:2010.00827*. 2020.

[44] F. Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.

[45] D. Zhou, L. Li, and Q. Gu. Neural contextual bandits with ucb-based exploration. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sect. 3, Sect. 4, and Sect. 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We do not see any potentially negative societal impact of this specific work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sect. 2, Sect. 3 and Sect. 4.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] We did not run experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A] We did not run experiments.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] We did not run experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] We did not run experiments.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See the reference section.
   (b) Did you mention the license of the assets? [N/A] We did not use any licenced material.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

This appendix contains, beyond the proof of all results contained in the main body (Section A.1 and Section A.2), the extension of our model selection results to the non-frozen NTK case (Section A.3). Section A.4 contains ancillary technical lemmas used throughout the proofs.

A.1 Proofs for Section 3

We first recall the following representation theorem (which is Lemma 5.1 in [45]). We give a proof sketch for completeness.

Lemma 1. *There exists a positive constant C such that for any $\delta \in (0,1)$, if*

$$m \geq CT^4 n^6 \log(2Tn/\delta)/\lambda_0^4$$

then with probability at least $1-\delta$ over the random initialization θ_0, there exists $\theta^* \in \mathbb{R}^p$ for which

$$h(x_{t,a}) = \langle g(x_{t,a};\theta_0), \theta^* - \theta_0 \rangle \qquad \text{and} \qquad \sqrt{m}\,\|\theta^* - \theta_0\|_2 \leq \sqrt{2} S_{T,n}(h) \tag{4}$$

for all $t \in [T]$, $a \in \mathcal{Y}$, and h.

Proof. Recall the rearrangement of $\{x_{t,a}\}_{t=1,\ldots,T,\, a=\pm 1}$ into $\{x^{(i)}\}_{i=1,\ldots,2T}$. We define the $p \times 2T$ matrix $G = \big[\phi(x^{(1)}), \ldots, \phi(x^{(2T)})\big]$. For $m = \Omega(T^4 n^6 \log(2Tn/\delta)/\lambda_0^4)$, we have $\|G^\top G - H\|_F \leq \lambda_0/2$ with probability at least $1-\delta$ over the random initialization over θ_0, which is based on a union bound over Theorem 3.1 in [2]. Since H on $\{x^{(i)}\}_{i=1,\ldots,2T}$ is positive definite with smallest eigenvalue λ_0, $G^\top G$ is also positive definite. Let the singular value decomposition of G be $G = PAQ^\top$, $P \in \mathbb{R}^{p\times 2T}$, $A \in \mathbb{R}^{2T\times 2T}$, $Q \in \mathbb{R}^{2T\times 2T}$, then A is also positive definite. We define

$$\theta^* = \theta_0 + PA^{-1}Q^\top \mathbf{h}/\sqrt{m}\,.$$

It is easy to see that θ^* satisfies (4), hence concluding the proof. □

Next we present a lemma relating the matrix Z_T with NTK matrix H.

Lemma 2. *There exists a positive constant C such that for any $\delta \in (0,1)$, if*

$$m \geq CT^6 n^6 \log(Tn/\delta)$$

then with probability at least $1-\delta$ over the random initialization θ_0 we have

$$\log\det Z_T \leq \log\det(I+H) + 1\,. \tag{5}$$

Proof. The proof is an adaptation of the proof of Lemma 5.4 in [45]. Let $G = (\phi(x^{(1)}, ..., \phi(x^{(2T)})) \in \mathbb{R}^{p\times 2T}$. We can write

$$\begin{aligned}
\log\det Z_T &= \log\det\left(I + \sum_{t=1}^{T} I_t \phi(x_{t,a_t})\phi(x_{t,a_t})^\top\right) \\
&\leq \log\det\left(I + \sum_{i=1}^{2T} \phi(x^{(i)})\phi(x^{(i)})^\top\right) \\
&= \log\det\big(I + GG^\top\big) \\
&= \log\det\big(I + G^\top G\big) \\
&= \log\det\big(I + H + (G^\top G - H)\big) \\
&\leq \log\det\big(I+H\big) + \langle (I+H)^{-1}, (G^\top G - H)\rangle_F \\
&\leq \log\det\big(I+H\big) + \|(I+H)^{-1}\|_F \|G^\top G - H\|_F \\
&\leq \log\det\big(I+H\big) + \sqrt{2T}\,\|G^\top G - H\|_F \\
&\leq \log\det(I+H) + 1\,.
\end{aligned}$$

In the above, the first inequality is obvious, the second inequality uses the fact that $\log\det(\cdot)$ is a concave function, the third one used Cauchy-Schwartz inequality, the fourth one comes from $\|(I+H)^{-1}\|_F \leq \|I\|_F = \sqrt{2T}$, and the last inequality uses Lemma B.1 in [45] along with our choice of m. □

The proofs of both Lemma 1 and Lemma 2 rely on controlling the size of $\|G^\top G - H\|_F$, which is small with high probability when m is large enough. Therefore, given

$$m \geq CT^4 \log(2Tn/\delta) n^6 \left(T^2 \vee 1/\lambda_0^4\right) ,$$

we have

$$\mathcal{E}_0 = \{\exists\, \theta^* \in \mathbb{R}^p \,:\, (4) \text{ and } (5) \text{ hold}\} , \tag{6}$$

holds with probability at least $1-\delta$ over random initialization of θ_0.

To take into account the random noise from the sequence of labels, we also define

$$\mathcal{E} = \{\exists\, \theta^* \in \mathbb{R}^p \,:\, \mathcal{E}_0 \text{ holds and } \theta^* \in \mathcal{C}_t \;\forall t > 0\} . \tag{7}$$

In order to make sense of the querying threshold B_t in Algorithm 1, we derive an upper and a lower bound for $U_{t,a} - h(x_{t,a})$ under $\mathcal{E}$.

As for the lower bound, simply notice that, by definition ,

$$U_{t,a} = \max_{\theta \in \mathcal{C}_{t-1}} \langle g(x_{t,a};\theta_0), \theta - \theta_0 \rangle \geq \langle g(x_{t,a};\theta_0), \theta^* - \theta_0 \rangle = h(x_{t,a}) . \tag{8}$$

To derive an upper bound, we can write

$$\begin{aligned}
U_{t,a} - h(x_{t,a}) &= \max_{\theta \in \mathcal{C}_{t-1}} \langle g(x_{t,a};\theta_0), \theta - \theta_0 \rangle - \langle g(x_{t,a};\theta_0), \theta^* - \theta_0 \rangle \\
&= \max_{\theta \in \mathcal{C}_{t-1}} \langle g(x_{t,a};\theta_0), \theta - \theta_{t-1} \rangle - \langle g(x_{t,a};\theta_0), \theta^* - \theta_{t-1} \rangle \\
&\leq \max_{\theta \in \mathcal{C}_{t-1}} \|g(x_{t,a};\theta_0)\|_{Z_{t-1}^{-1}} \Big(\|\theta - \theta_{t-1}\|_{Z_{t-1}} + \|\theta^* - \theta_{t-1}\|_{Z_{t-1}} \Big) \\
&\leq 2\gamma_{t-1} \|\phi(x_{t,a})\|_{Z_{t-1}^{-1}} ,
\end{aligned} \tag{9}$$

where in the last inequality we used the definition of $\mathcal{C}_{t-1}$ and the assumption that $\theta^* \in \mathcal{C}_{t-1}$. A proof of this assumption is contained in the below lemma, which follows from standard arguments.

Lemma 3. *Let the input parameter S in Algorithm 1 be such that $\sqrt{2} S_{T,n}(h) \leq S$, then under event $\mathcal{E}_0$ for any $\delta > 0$, with probability at least $1-\delta$ over the random noises we have*

$$\|\theta^* - \theta_t\|_{Z_t} \leq \gamma_t / \sqrt{m}$$

for all $t \geq 0$ simultaneously, i.e., $\theta^* \in \mathcal{C}_t$ with high probability simultaneously for all $t \geq 0$.

Proof. We essentially follow the proof of Theorem 2 in [1] (see also the proof of Lemma 5.2 in [45]).

We have $\ell_t = 1 - h(x_{t,a_t}) - \xi_t$, where $\xi_t = 1 - \ell_t - h(x_{t,a_t})$ is a sub-Gaussian random variable. Hence, setting $\boldsymbol{\xi}_t = (I_1\xi_1, ..., I_t\xi_t)^\top$, $X_t = (I_1\phi(x_{1,a_1}), ..., I_t\phi(x_{t,a_t}))^\top$, and $Y_t = (I_1(1-\ell_1), ..., I_t(1-\ell_t))^\top$, we can write

$$Z_t = X_t^\top X_t + I, \qquad b_t = X_t^\top Y_t$$

Plug them into the definition of θ_t gives

$$\begin{aligned}
\theta_t - \theta_0 &= Z_t^{-1} b_t / \sqrt{m} \\
&= (X_t^\top X_t + I)^{-1} X_t^\top (\sqrt{m} X_t (\theta^* - \theta_0) + \boldsymbol{\xi}_t)/\sqrt{m} \\
&= (X_t^\top X_t + I)^{-1} X_t^\top \boldsymbol{\xi}_t / \sqrt{m} + \theta^* - \theta_0 - (X_t^\top X_t + I)^{-1} (\theta^* - \theta_0) ,
\end{aligned}$$

where in the first equality we used definition of ξ_t and Lemma 1. Now, for any $x \in \mathbb{R}^p$, we get

$$x^\top (\theta_t - \theta^*) = \langle x, X_t^\top \boldsymbol{\xi}_t \rangle_{Z_t^{-1}} / \sqrt{m} - \langle x, \theta^* - \theta_0 \rangle_{Z_t^{-1}} ,$$

hence

$$\begin{aligned}
|x^\top (\theta_t - \theta^*)| &\leq \|x\|_{Z_t^{-1}} \Big(\|X_t^\top \boldsymbol{\xi}_t\|_{Z_t^{-1}} / \sqrt{m} + \|\theta^* - \theta_0\|_{Z_t^{-1}} \Big) \\
&\leq \|x\|_{Z_t^{-1}} \Big(\|X_t^\top \boldsymbol{\xi}_t\|_{Z_t^{-1}} / \sqrt{m} + \|\theta^* - \theta_0\|_2 \Big) ,
\end{aligned}$$

where the first inequality derives from the Cauchy-Schwartz inequality and the second from the fact that the smallest eigenvalue of Z_t is at least 1. Then, by Theorem 1 in [1], for any δ with probability at least $1-\delta$ over the random noises

$$\|X_t^\top \boldsymbol{\xi}_t\|_{Z_t^{-1}} \le \sqrt{\log\left(\frac{\det(Z_t)}{\delta^2}\right)} \,.$$

Therefore, when $\mathcal{E}_0$ holds, we have for all $t>0$, with probability at least $1-\delta$,

$$|x^\top(\theta_t-\theta^*)| \le \|x\|_{Z_t^{-1}} \left(\sqrt{\log\left(\frac{\det(Z_t)}{\delta^2}\right)/m} + \sqrt{2}S_{T,n}(h)/\sqrt{m}\right) \,.$$

Plugging in $x = Z_t(\theta_t-\theta^*)$ and using $\sqrt{2}S_{T,n}(h) \le S$, we obtain

$$\|\theta^*-\theta_t\|_{Z_t} \le \sqrt{\log\left(\frac{\det(Z_t)}{\delta^2}\right)/m} + S/\sqrt{m} = \gamma_t/\sqrt{m} \,,$$

as claimed. □

Combining Lemma 1, 2 and 3 we confirm that $\mathcal{E}$ is a high probability event.

Lemma 4. *There exists a constant C such that if $m \ge CT^4\log(2Tn/\delta)n^6\left(T^2 \vee 1/\lambda_0^4\right)$ and $\sqrt{2}S_{T,n}(h) \le S$, then*

$$\mathbb{P}(\mathcal{E}) \ge 1-2\delta \,. \tag{10}$$

Proof. Lemma 1 and 2 imply that $\mathbb{P}(\mathcal{E}_0) \ge 1-\delta$ when $m \ge CT^4\log(2Tn/\delta)n^6\left(T^2 \vee 1/\lambda_0^4\right)$. Lemma 3 implies that when $\sqrt{2}S_{T,n}(h) \le S$, $\mathbb{P}(\theta^* \in \mathcal{C}_t \ \forall t>0 \mid \mathcal{E}_0) \ge 1-\delta$. Therefore,

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}(\theta^* \in \mathcal{C}_t \ \forall t>0 \mid \mathcal{E}_0)\mathbb{P}(\mathcal{E}_0) \ge (1-\delta)^2 \ge 1-2\delta \,.$$

□

Lemma 5. *For any $b>0$ we have*

$$\sum_{t=1}^T b \wedge I_t B_t^2 \le 8\left(\log\det Z_T + 2\log(1/\delta) + S^2 + \frac{b}{8}\right)\log\det Z_T \,. \tag{11}$$

Proof. By definition of B_t and the fact that γ_t is increasing, we have

$$\sum_{t=1}^T b \wedge I_t B_t^2 \le 4\gamma_T^2 \sum_{t=1}^T \frac{b}{4\gamma_T^2} \wedge I_t\|\phi(x_{t,a_t})\|^2_{Z_{t-1}^{-1}} \le (b+4\gamma_T^2)\log\det Z_T \,,$$

where the second inequality is from Lemma 24. Using the definition of γ_T and the inequality $(a+b)^2 \le 2a^2+2b^2$ we obtain

$$\gamma_T^2 \le 2\log\det Z_T + 4\log(1/\delta) + 2S^2 \,.$$

Plugging this in we get (11). □

Let us now introduce the short-hand notation

$$\widehat{\Delta}_t = U_{t,a_t} - 1/2 \,, \qquad \Delta_t = h(x_{t,a_t}) - 1/2 \,, \qquad T_\epsilon = \sum_{t=1}^T \mathbb{1}\{\Delta_t^2 \le \epsilon^2\} \,,$$

for some $\epsilon \in (0, \frac{1}{2})$. Combined with (8) and (9), we have the following statement about $\widehat{\Delta}_t$ and Δ_t.

Lemma 6. *Under event $\mathcal{E}$, $0 \le \widehat{\Delta}_t - \Delta_t \le B_t$ and $0 \le \widehat{\Delta}_t$ hold for all t, where B_t is the querying threshold in Algorithm 1, i.e.,*

$$B_t = 2\gamma_{t-1}\|\phi(x_{t,a_t})\|_{Z_{t-1}^{-1}} \,.$$

Proof. Recalling that (8) and (9) implies that for $a \in \mathcal{Y}$

$$0 \le U_{t,a} - h(x_{t,a}) \le B_t \ .$$

Specifically when $a = a_t$,

$$0 \le \widehat{\Delta}_t - \Delta_t \le B_t \ .$$

Also using (8) we have $U_{t,1} + U_{t,-1} \ge h(x_{t,1}) + h(x_{t,-1}) = 1$. Hence, by definition of a_t, $U_{t,a_t} \ge 1/2$, i.e., $\widehat{\Delta}_t \ge 0$. □

The following lemma bounds the label complexity N_T of Algorithm 1 under event $\mathcal{E}$. Notice that, as stated, the bound does not depend on any specific properties of the marginal distribution $\mathcal{D}_{\mathcal{X}}$.

Lemma 7. *Under event $\mathcal{E}$, for any $\epsilon \in (0, 1/2)$ we have*

$$\begin{aligned} N_T &\le T_\epsilon + \frac{8}{\epsilon^2}(\log\det Z_T + 2\log(1/\delta) + S^2 + \frac{1}{32})\log\det Z_T \\ &= O\left(T_\epsilon + \frac{1}{\epsilon^2}\left(\log\det(I+H) + \log(1/\delta) + S^2\right)\log\det(I+H)\right) \ . \end{aligned}$$

Proof. We adapt the proof of Lemma 6 in [16]. Assume $\mathcal{E}$ holds. Since $0 \le \widehat{\Delta}_t - \Delta_t \le B_t$ and $\widehat{\Delta}_t \ge 0$ by Lemma 6, $\hat{\Delta}_t \le B_t$ implies $|\Delta_t| \le B_t$. We can write

$$\begin{aligned} I_t &= I_t \mathbb{1}\Big\{\hat{\Delta}_t \le B_t\Big\} \\ &\le I_t \mathbb{1}\Big\{\hat{\Delta}_t \le B_t, B_t \ge \epsilon\Big\} + I_t \mathbb{1}\Big\{\widehat{\Delta}_t \le B_t, B_t < \epsilon\Big\} \\ &\le \frac{I_t B_t^2}{\epsilon^2} \wedge 1 + \mathbb{1}\big\{\Delta_t^2 \le \epsilon^2\big\} \ . \end{aligned}$$

For the first term, summing over t yields

$$\begin{aligned} \frac{1}{\epsilon^2}\sum_{t=1}^T I_t B_t^2 \wedge \epsilon^2 &\le \frac{1}{\epsilon^2}\sum_{t=1}^T I_t B_t^2 \wedge \frac{1}{4} \\ &\le \frac{8}{\epsilon^2}\left(\log\det Z_T + 2\log(1/\delta) + S^2 + \frac{1}{32}\right)\log\det Z_T \\ &= O\left(\frac{1}{\epsilon^2}\left(\log\det(I+H) + \log(1/\delta) + S^2\right)\log\det(I+H)\right) \ , \end{aligned}$$

where the second bound follows from Lemma 5, and the last bound holds under event $\mathcal{E}$. □

The next lemma shows that on rounds where Algorithm 1 does not issue a query, we are confident that our prediction a_t suffers no regret.

Lemma 8. *Under event $\mathcal{E}$, for the rounds t such that $I_t = 0$, we have $a_t = a_t^*$, that is, Algorithm 1 suffers no regret.*

Proof. We apply Lemma 6, when $I_t = 0$ this yields $\widehat{\Delta}_t > B_t$. As a consequence of the condition $\widehat{\Delta}_t - \Delta_t \le B_t$, we get $\Delta_t > 0$, which in turn entails $a_t = a_t^*$. □

The next lemma establishes an upper bound on the cumulative regret R_T in the same style as in Lemma 7.

Lemma 9. *Under event $\mathcal{E}$, for any $\epsilon \in (0, 1/2)$ we have*

$$\begin{aligned} R_T &\le 2\epsilon T_\epsilon + \frac{16}{\epsilon}\left(\log\det Z_T + 2\log(1/\delta) + S^2 + \frac{1}{16}\right)\log\det Z_T \\ &= O\left(\epsilon T_\epsilon + \frac{1}{\epsilon}\left(\log\det(I+H) + \log(1/\delta) + S^2\right)\log\det(I+H)\right) \ . \end{aligned}$$

Proof. By virtue of Lemma 8, we can restrict with high probability to the rounds t on which $I_t = 1$. We have

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} I_t\big(h(x_{t,a_t^*}) - h(x_{t,a_t})\big) \\
&= \sum_{t=1}^{T} I_t\big(h(x_{t,a_t^*}) - h(x_{t,a_t})\big)\, \mathbb{1}\{a_t \neq a_t^*\} \\
&\leq \sum_{t=1}^{T} I_t\big|h(x_{t,1}) - h(x_{t,-1})\big|\, \mathbb{1}\{a_t \neq a_t^*\} \\
&= 2\sum_{t=1}^{T} I_t|\Delta_t| \\
&= 2\sum_{t=1}^{T} I_t|\Delta_t|\, \mathbb{1}\{|\Delta_t| > \epsilon\} + 2\sum_{t=1}^{T} I_t|\Delta_t|\, \mathbb{1}\{|\Delta_t| \leq \epsilon\}\,.
\end{aligned}$$

The second sum is clearly upper bounded by $2\epsilon T_\epsilon$. As for the first sum, notice that Lemma 6 along with $I_t = 1$ implies $|\Delta_t| \leq B_t$ under event $\mathcal{E}$. Therefore

$$\begin{aligned}
2\sum_{t=1}^{T} I_t|\Delta_t|\, \mathbb{1}\{|\Delta_t| > \epsilon\} &\leq \frac{2}{\epsilon}\sum_{t=1}^{T} I_t\Delta_t^2 \wedge \epsilon \\
&\leq \frac{2}{\epsilon}\sum_{t=1}^{T} I_t B_t^2 \wedge \frac{1}{2} \\
&\leq \frac{16}{\epsilon}\left(\log\det Z_T + 2\log(1/\delta) + S^2 + \frac{1}{16}\right)\log\det Z_T \\
&= O\left(\frac{1}{\epsilon}\left(\log\det(I+H) + \log(1/\delta) + S^2\right)\,\log\det(I+H)\right)\,.
\end{aligned}$$

The third bound follows from Lemma 5, while the last bound holds under event $\mathcal{E}$. □

At this point, we leverage the fact that $x_1, ..., x_T$ are generated in an i.i.d. fashion according to a marginal distribution $\mathcal{D}_\mathcal{X}$ satisfying the low-noise assumption with exponent α recalled in Section 3. A direct application of Lemma 23 (Appendix A.4) gives, with probability at least $1-\delta$,

$$T_\epsilon \leq 3T\epsilon^\alpha + O\left(\log\frac{\log T}{\delta}\right)\,,$$

simultaneously over ϵ. Using the above bound on T_ϵ back into both Lemma 7 and Lemma 9 and optimizing over ϵ in the two bounds separately yields the following result, which is presented in the main body as Theorem 1.

Theorem 3. *Let Algorithm 1 be run with parameters δ, S, m, and n on an i.i.d. sample $(x_1, y_1), \ldots, (x_T, y_T) \sim \mathcal{D}$, where the marginal distribution $\mathcal{D}_\mathcal{X}$ fulfills the low-noise condition with exponent $\alpha \geq 0$ w.r.t. a function h that satisfies (1) and such that $\sqrt{2}S_{T,n}(h) \leq S$ for all $\{x_i\}_{i=1}^T$. Also assume $m \geq CT^4\log(2Tn/\delta)n^6\left(T^2 \vee 1/\lambda_0^4\right)$ where C is the constant in Lemma 1 and Lemma 2. Then with probability at least $1-\delta$ the cumulative regret R_T and the total number of queries N_T are simultaneously upper bounded as follows:*

$$\begin{aligned}
R_T &= O\left(L_H^{\frac{\alpha+1}{\alpha+2}}\left(L_H + \log(1/\delta) + S^2\right)^{\frac{\alpha+1}{\alpha+2}} T^{\frac{1}{\alpha+2}} + \log(\log T/\delta)\right) \\
N_T &= O\left(L_H^{\frac{\alpha}{\alpha+2}}\left(L_H + \log(1/\delta) + S^2\right)^{\frac{\alpha}{\alpha+2}} T^{\frac{2}{\alpha+2}} + \log(\log T/\delta)\right)\,,
\end{aligned}$$

where $L_H = \log\det(I+H)$, and H is the NTK matrix of depth n over the set of points $\{x_{t,a}\}_{t=1,\ldots,T,\, a=\pm 1}$.

A.2 Proofs for Section 4

Additional notation. In this section, we add subscript "i" to the relevant quantities occurring in the proof when these quantities refer to the i-th base learner. For instance, we write $Z_{t,i}$ to denote the covariance matrix updated within the i-th base learner, $B_{t,i} = B_{t,i}(S_i) = 2\gamma_{t-1,i}\|\phi(x_{t,a_t})\|_{Z_{t-1,i}^{-1}}$, with $\gamma_{t-1,i} = \sqrt{\log\det Z_{t-1,i} + 2\log(1/\delta)} + S_i$, and $\mathcal{C}_{t,i}$ to denote the confidence ellipsoid maintained by the i-th base learner.

For convenience, we also introduce the function

$$d(S,\delta) = (\log\det(I+H)+1)(\log\det(I+H) + \frac{17}{16} + 2\log(M/\delta) + S^2)\,. \tag{12}$$

The above is a high probability upper bound on $(\frac{1}{16} + \frac{1}{2}\gamma_{T,i}^2)\log\det Z_{T,i}$ (holding for all i), which in turn upper bounds $\frac{1}{8}\sum_{t=1}^{T} I_{t,i}B_{t,i}^2 \wedge \frac{1}{2}$.

By the assumption in Theorem 2, we know that there is a learner $i^\star = \langle i_1^\star, i_2^\star\rangle \in \mathcal{M}_1$ such that its parameters $S_{i_1^\star}$ and $d_{i_2^\star}$ satisfy

$$\sqrt{2}S_{T,n}(h) \le S_{i_1^\star} \le 2\sqrt{2}S_{T,n}(h) \tag{13}$$

$$d(S_{T,n}(h),\delta) \le d(S_{i_1^\star},\delta) \le d_{i_2^\star} \le 2d(S_{i_1^\star},\delta) \le 8d(S_{T,n}(h),\delta)\,. \tag{14}$$

Throughout the proof we will refer to a specific learner that satisfies these conditions by $i^\star$. Moreover, we denote by $\mathcal{E}_i$ the event where the conditions of the event in Eq. (7) and the event in Lemma 2 hold for base learner i. In $\mathcal{E}_i$, we call i well-specified.

Let $R(\mathcal{T})$ and $N(\mathcal{T})$ denote cumulative regret R and number of requested labels N when restricted to subset $\mathcal{T} \subseteq [T]$. Then the regret and label complexity analyses of Algorithm 1 in Section A.1 directly imply the following regret and label complexity bounds of a well-specified base learner i during the execution of Algorithm 2.

Lemma 10 (Regret and label complexity of a well-specified base learner)**.** *Let $i \in \mathcal{M}_1$ be any base learner. In event $\mathcal{E}_i$ (when i is well-specified), the following regret and label complexity bound holds for any $0 < \epsilon < \frac{1}{2}$ and $t \in [T]$:*

$$R(\mathcal{T}_{t,i}) \le 2\sum_{k\in\mathcal{T}_{t,i}} I_{k,i}B_{k,i} \wedge \frac{1}{2} \;\le\; \frac{16}{\epsilon}\,d(S_{i_1},\delta) + 2\epsilon|\mathcal{T}_{t,i}^{\epsilon}|$$

$$N(\mathcal{T}_{t,i}) \le |\mathcal{T}_{t,i}^{\epsilon}| + \frac{1}{\epsilon^2}\sum_{k\in\mathcal{T}_{t,i}} I_{k,i}B_{k,i}^2 \wedge \frac{1}{4} \;\le\; \frac{8}{\epsilon^2}\,d(S_{i_1},\delta) + |\mathcal{T}_{t,i}^{\epsilon}|\,,$$

where $\mathcal{T}_{t,i}^{\epsilon} = \{k \in [t] \colon i_k = i,\ |\Delta_k| \le \epsilon\}$. Furthermore, in rounds $t \in \mathcal{T}_{t,i}$ where the label is not queried ($I_{t,i} = 0$), the regret is 0.

Proof. This follows directly from the analysis of Algorithm 1 in the previous section. □

Equipped with these two properties of well-specified base learners, we can first show that with high probability, Algorithm 2 will never eliminate a well-specified learner, and subsequently analyze the label complexity and cumulative regret of Algorithm 2.

Lemma 11. *Let $i = \langle i_1, i_2\rangle \in \mathcal{M}_1$ be a base learner with $d_{i_2} \ge d(S_{i_1},\delta)$. Assume $\gamma \le \alpha$ and consider event $\bigcap_{j\colon j\ge i_1}\mathcal{E}_j$. Then, under that event, with probability at least $1 - M\delta$ Algorithm 2 never eliminates base learner i.*

Proof. We show the statement for each of the four mis-specification tests in turn:

- **Disagreement test:** Consider a round t and any learner $j = \langle j_1, j_2\rangle$ with $S_{j_1} \ge S_{i_1}$ and $I_{t,i} = I_{t,j} = 0$. By assumption, $\mathcal{E}_i \cap \mathcal{E}_j$ holds. Since i did not ask for the label, this implies that $|\Delta_t| > 0$ (since in rounds with no margin $|\Delta_t| = 0$, a learner always asks for the label). Further, by Lemma 10, the prediction of i and j has no regret in round t. Thus, i and j need to make the same prediction and the test does not trigger.

- **Observed regret test:** Consider a round t and any $j \in \mathcal{M}_t$. Then, by virtue of Lemma 21 (Appendix A.4), the left-hand side of the observed regret test for pair (i, j) is upper-bounded with probability at east $1 - \delta$ as

$$\begin{aligned}
&\sum_{k \in \mathcal{V}_{t,i,j}} \left(\mathbb{1}\{a_{k,i} \neq y_k\} - \mathbb{1}\{a_{k,j} \neq y_k\} \right) \\
&\qquad \leq \sum_{k \in \mathcal{V}_{t,i,j}} \left(h(x_{k,a_{k,j}}) - h(x_{k,a_{k,i}}) \right) + 0.72 \sqrt{|\mathcal{V}_{t,i,j}| L(|\mathcal{V}_{t,i,j}|, \delta)} \\
&\qquad \leq \sum_{k \in \mathcal{V}_{t,i,j}} \left(h(x_{k,a_k^\star}) - h(x_{k,a_{k,i}}) \right) + 0.72 \sqrt{|\mathcal{V}_{t,i,j}| L(|\mathcal{V}_{t,i,j}|, \delta)} \\
&\qquad = R(\mathcal{V}_{t,i,j}) + 0.72 \sqrt{|\mathcal{V}_{t,i,j}| L(|\mathcal{V}_{t,i,j}|, \delta)} \,,
\end{aligned}$$

  where the second inequality follows from the definition of the best prediction a_k^* for round k. Finally, in event $\mathcal{E}_i$ the regret of i in rounds $\mathcal{V}_{t,i,j}$ is bounded by Lemma 10 as

$$R(\mathcal{V}_{t,i,j}) \leq \sum_{k \in \mathcal{V}_{t,i,j}} 1 \wedge B_{k,i} \,.$$

  Therefore, this test does not trigger for pair (i, j) in round t. By a union bound, this happens with probability at least $1 - M\delta$.

- **Label complexity test:** By Lemma 10, the number of labels requested by i up to round t is at most

$$\sum_{k \in \mathcal{T}_{t,i}} I_{k,i} \leq \inf_{\epsilon \in (0, 1/2]} |\mathcal{T}_{t,i}^\epsilon| + \frac{1}{\epsilon^2} \sum_{k \in \mathcal{T}_{t,i}} I_{k,i} B_{k,i}^2 \wedge \frac{1}{4} \,.$$

  We now use Lemma 23 (Appendix A.4) to upper-bound $|\mathcal{T}_{t,i}^\epsilon|$ simultaneously for all ϵ as

$$|\mathcal{T}_{t,i}^\epsilon| \leq 3\epsilon^\gamma |\mathcal{T}_{t,i}| + 2L(|\mathcal{T}_{t,i}|, \delta / \log_2(12t)) \,.$$

  By plugging this expression into the previous bound (and taking a union bound over i) we show that the label complexity test is not triggered.

- d_i **test:** Using the assumption that $\mathcal{E}_i$ holds and Lemma 5, we can bound the left-hand side of the test as

$$\begin{aligned}
\sum_{k \in \mathcal{T}_{t,i}} \left(\frac{1}{2} \wedge I_{k,i} B_{k,i}^2 \right) &\leq 8(\log\det Z_{t,i} + 2\log(1/\delta) + S_{i_1}^2 + 1/16) \log\det Z_{t,i} \\
&\leq 8(\log\det(H + I) + 2\log(1/\delta) + S_{i_1}^2 + 17/16)(\log\det(H + I) + 1) \\
&= 8d(S_{i_1}, \delta)
\end{aligned}$$

  and by the assumption that $d_{i_2} \geq d(S_{i_1}, \delta)$, learner i is not be eliminated by this test.

This concludes the proof. □

A.2.1 Label Complexity Analysis

Lemma 12 (Label complexity of Algorithm 2)**.** *In event* $\bigcap_{i=\langle i_1, i_2 \rangle \in \mathcal{M}_1 \,:\, i_1 \geq i_1^\star} \mathcal{E}_i$*, Algorithm 2 queries with probability at least* $1 - M\delta$

$$N(T) = O\left(\sum_{i=\langle i_1, i_2 \rangle \in \mathcal{M}_1} \left(\frac{d_{i_2}}{\epsilon^2} + \epsilon^\gamma T \left(1 \wedge \frac{d(S_{T,n}(h), \delta)}{d_{i_2}} \right)^{\gamma+1} \right) + ML(T, \delta / \log T) \right)$$

labels.

Proof. We can decompose the total number of label requests as

$$N(T) = \sum_{t=1}^{T} I_{t,i_t} = \sum_{i=1}^{M} \sum_{t \in \mathcal{T}_{T,i}} I_{t,i} = \sum_{i \in \mathcal{M}_1} N(\mathcal{T}_{T,i}) \,.$$

Since each learner i satisfied the label complexity test except possibly for the round where it was eliminated, we have

$$\begin{aligned} N(\mathcal{T}_{T,i}) &= O\left(\inf_{\epsilon \in (0,1/2)} \left(\epsilon^{\gamma} |\mathcal{T}_{T,i}| + \frac{1}{\epsilon^2} \sum_{k \in \mathcal{T}_{t,i}} I_{k,i} B_{k,i}^2 \wedge \frac{1}{4} \right) + L(|\mathcal{T}_{T,i}|, \delta / \log t) \right) \\ &= O\left(\inf_{\epsilon \in (0,1/2)} \left(\epsilon^{\gamma} \sum_{k \in [T]} p_{k,i} + \frac{1}{\epsilon^2} \sum_{k \in \mathcal{T}_{t,i}} I_{k,i} B_{k,i}^2 \wedge \frac{1}{4} \right) + L(T, \delta / \log T) \right) \\ &= O\left(\inf_{\epsilon \in (0,1/2)} \left(\epsilon^{\gamma} \sum_{k \in [T]} p_{k,i} + \frac{d_{i_2}}{\epsilon^2} \right) + L(T, \delta / \log T) \right) \,, \end{aligned} \tag{15}$$

where the second inequality holds with probability at least $1 - \delta$ by Lemma 22 and the final inequality holds by the d_i test. We now bound $\sum_{k \in [T]} p_{k,i}$ as

$$\sum_{k \in [T]} p_{k,i} \leq T(1 \wedge d_{i_2}^{-(\gamma+1)} d_{i_2^\star}^{\gamma+1}) \leq T d_{i_2}^{-(\gamma+1)} (8 d(S_{T,n}(h), \delta))^{\gamma+1} \wedge T$$

where we used that by Lemma 11 learner $i^\star$ never gets eliminated in the considered event.

□

A.2.2 Regret Analysis

To bound the overall cumulative regret of Algorithm 2, we decompose the rounds $[T]$ into the following three disjoint sets of rounds

$$[T] = \mathcal{R}_{i^\star} \dot\cup \, \mathcal{U}_{i^\star} \dot\cup \, \mathcal{O}_{i^\star}, \tag{16}$$

where

- $\mathcal{R}_{i^\star} = \{t \in [T] \colon I_{t,i^\star} = 1\}$ are the rounds where $i^\star$ requests a label,
- $\mathcal{U}_{i^\star} = \{t \in [T] \colon I_{t,i^\star} = 0, I_{t,i_t} = 0\}$ are the rounds where $i^\star$ does not request the label and the label was not observed,
- $O_{i^\star} = \{t \in [T] \colon I_{t,i^\star} = 0, I_{t,i_t} = 1\}$ are the rounds where $i^\star$ does not request the label and the label was observed.

In the following three lemmas, we bound the regret in these sets of rounds separately.

Lemma 13 (Regret in rounds where $i^\star$ requests)**.** *In event $\bigcap_{i=\langle i_1, i_2 \rangle \in \mathcal{M}_1 \colon i_1 \geq i_1^\star} \mathcal{E}_i$, the regret in rounds where $i^\star = \langle i_1^\star, i_2^\star \rangle$ would request the label is bounded with probability at least $1 - \delta$ for all $\epsilon \in (0, 1/2)$ as*

$$R(\mathcal{R}_{i^\star}) = O\left(\frac{M}{\epsilon} 2^{\gamma+1} d(S_{i_1^\star}, \delta)^{\gamma+2} + \frac{M}{\epsilon} 2^{\gamma+1} d(S_{i_1^\star}, \delta)^{\gamma+1} L(T, \delta) + \epsilon T_\epsilon \right) \,. \tag{17}$$

Proof. In any round, the largest instantaneous regret possible is $2|h(x_{t,1}) - 1/2| = 2|h(x_{t,-1}) - 1/2| = 2|\Delta_{t,i^\star}|$, no matter whether the prediction of $i^\star$ was followed or not. Thus, the regret in rounds $\mathcal{R}_{i^\star}$ can be bounded as

$$R(\mathcal{R}_{i^\star}) \leq 2 \sum_{t \in \mathcal{R}_{i^\star}} |\Delta_{t,i^\star}| = 2 \sum_{t \in \mathcal{R}_{i^\star}} \mathbb{1}\{|\Delta_{t,i^\star}| > \epsilon\} |\Delta_{t,i^\star}| + 2\epsilon |\mathcal{R}_{i^\star}^\epsilon|,$$

for any $\epsilon \in (0, 1/2)$ where $\mathcal{R}_{i^\star}^\epsilon = \{t \in \mathcal{R}_{i^\star} \colon |\Delta_t| \leq \epsilon\}$.

On rounds $\mathcal{R}_{i^\star}$, learner $i^\star$ wants to query the label which means $\widehat{\Delta}_{t,i^\star} \le B_{t,i^\star}$. Moreover in $\mathcal{E}_{i^\star}$, the conditions $0 \le \widehat{\Delta}_{t,i^\star} - \widehat{\Delta}_{t,i^\star} \le B_{t,i^\star}$ and $0 \le \widehat{\Delta}_{t,i^\star}$ hold. Combining both inequalities gives $|\Delta_{t,i^\star}| \le B_{t,i^\star}$ and we can further bound the display above as

$$\begin{aligned}
R(\mathcal{R}_{i^\star}) &\le \sum_{t\in\mathcal{R}_{i^\star}} \mathbb{1}\{|\Delta_{t,i^\star}| > \epsilon\}(1 \wedge 2B_{t,i^\star}) + 2\epsilon|\mathcal{R}^\epsilon_{i^\star}| \\
&\le \sum_{t\in\mathcal{R}_{i^\star}} \mathbb{1}\{|\Delta_{t,i^\star}| > \epsilon\}\left(1 \wedge \frac{2B^2_{t,i^\star}}{\epsilon}\right) + 2\epsilon|\mathcal{R}^\epsilon_{i^\star}| \\
&\le \frac{2}{\epsilon}\sum_{t\in\mathcal{R}_{i^\star}} \left(\frac{\epsilon}{2} \wedge B^2_{t,i^\star}\right) + 2\epsilon|\mathcal{R}^\epsilon_{i^\star}| \,.
\end{aligned}$$

To bound the remaining sum, we appeal to the randomized potential lemma in Lemma 25. We denote $\underline{p}^\star = \min_{k\in[T]} p_{k,i^\star}$ the smallest probability of $i^\star$ in any round. Then Lemma 25 gives with probability at least $1-\delta$

$$\begin{aligned}
\sum_{t\in\mathcal{R}_{i^\star}} \left(\frac{\epsilon}{2} \wedge B^2_{t,i^\star}\right) &\le \sum_{t\in\mathcal{R}_{i^\star}} \left(\frac{1}{4} \wedge B^2_{t,i^\star}\right) \le 4\gamma^2_{T,i^\star} \sum_{t\in\mathcal{R}_{i^\star}} \left(\frac{1}{16\gamma^2_{T,i^\star}} \wedge \|\phi(x_{t,a_{t,i^\star}})\|^2_{Z^{-1}_{t-1,i^\star}}\right) \\
&\le 4\gamma^2_{T,i^\star}\left(1 + \frac{3}{16\underline{p}^\star\gamma^2_{T,i^\star}}L(T,\delta)\right) + \frac{8\gamma^2_{T,i^\star}}{\underline{p}^\star}(1 + \frac{1}{16\gamma^2_{T,i^\star}})\log\det Z_{T,i^\star} \\
&\le \frac{12\gamma^2_{T,i^\star} + \frac{1}{2}}{\underline{p}^\star}\log\det Z_{T,i^\star} + \frac{3}{4\underline{p}^\star}L(T,\delta) \,,
\end{aligned}$$

because $\gamma_{t,i^\star}$ is non-decreasing in T. Plugging this back into the previous display yields

$$\begin{aligned}
R(\mathcal{R}_{i^\star}) &\le 24\frac{\gamma^2_{T,i^\star} + \frac{1}{24}}{\epsilon\underline{p}^\star}\log\det Z_{T,i^\star} + \frac{3}{2\epsilon\underline{p}^\star}L(T,\delta) + 2\epsilon|\mathcal{R}^\epsilon_{i^\star}| \\
&\le 48\frac{d(S_{i^\star_1},\delta)}{\epsilon\underline{p}^\star} + \frac{3}{2\epsilon\underline{p}^\star}L(T,\delta) + 2\epsilon T_\epsilon \,.
\end{aligned}$$

Now, Lemma 11 ensures that $i^\star$ never gets eliminated in the considered event. Therefore

$$\frac{1}{\underline{p}^\star} \le \frac{\sum_{i\in\mathcal{M}_1} d^{-(\gamma+1)}_{i_2}}{d^{-(\gamma+1)}_{i^\star_2}} = d^{\gamma+1}_{i^\star_2} M \le M(2d(S_{i^\star_1},\delta))^{\gamma+1} \,,$$

where the last inequality follows from Eq. (13). Plugging this back into the previous display yields

$$R(\mathcal{R}_{i^\star}) \le \frac{48M}{\epsilon}2^{\gamma+1}d(S_{i^\star_1},\delta)^{\gamma+2} + \frac{3M}{2\epsilon}2^{\gamma+1}d(S_{i^\star_1},\delta)^{\gamma+1}L(T,\delta) + 2\epsilon T_\epsilon \,,$$

as claimed. □

Lemma 14 (Regret in unobserved rounds where $i^\star$ does not request)**.** *In event $\mathcal{E}_{i^\star}$,*

$$R(\mathcal{U}_{i^\star}) \le M \,. \tag{18}$$

Proof. If $i^\star$ is not requesting the label then $i^\star$ predicts the label as a^*_t. From the disagreement test i_t will predict the same label as $i^\star$ so there should be no regret, except when a learner gets eliminated. Since there are at most M learners and the regret per round is at most 1, the total regret on rounds $\mathcal{U}_{i^\star}$ can at most be M. □

Lemma 15 (Regret in observed rounds where $i^\star$ does not request)**.** *In event $\bigcap_{i=\langle i_1,i_2\rangle\in\mathcal{M}_1 \,:\, i_1\ge i^\star_1} \mathcal{E}_i$, the regret in rounds where $i^\star$ does not request the label, but the label was still observed is bounded as*

$$\begin{aligned}
&R(\mathcal{O}_{i^\star}) \\
&= O\left(\sum_{i=\langle i_1,i_2\rangle\in\mathcal{M}_1} \inf_{\epsilon\in(0,1/2)}\left(\frac{d_{i_2}}{\epsilon} + T\left(\frac{\epsilon\, d(S_{T,n}(h),\delta)}{d_{i_2}}\right)^{\gamma+1} + \frac{L(T,\delta)}{\epsilon}\right) + ML(T,\delta/\log T)\right) \,.
\end{aligned}$$

Proof. Note that we can decompose the regret in those rounds as

$$R(\mathcal{O}_{i^\star}) = \sum_{i \neq i_*} R(\mathcal{V}_{T,i,i^\star})$$

since no regret occurs if the played action agrees with the action proposed by $i^\star$ which did not request a label and in $\mathcal{E}_{i^\star}$ does not incur any regret in such rounds. We bound $R(\mathcal{V}_{T,i,i^\star})$ by using the fact that in all but at most one of those rounds both the observed regret test and the d_i test did not trigger. This gives

$$\sum_{k \in \mathcal{V}_{T,i,i^\star}} (\mathbb{1}\{a_{k,i} \neq y_k\} - \mathbb{1}\{a_{k,i^\star} \neq y_k\}) \leq \sum_{k \in \mathcal{V}_{T,i,i^\star}} 1 \wedge B_{k,i} + 1.45\sqrt{|\mathcal{V}_{T,i,i^\star}| L(|\mathcal{V}_{T,i,i^\star}|, \delta)} + 1 \,.$$

We now apply the concentration argument in Lemma 21 to bound the LHS from below as

$$\begin{aligned}
&\sum_{k \in \mathcal{V}_{T,i,i^\star}} (\mathbb{1}\{a_{k,i} \neq y_k\} - \mathbb{1}\{a_{k,i^\star} \neq y_k\}) \\
&\geq \sum_{k \in \mathcal{V}_{T,i,i^\star}} (h(x_{k,a_{k,i^\star}}) - h(x_{k,a_{k,i}})) - 0.72\sqrt{|\mathcal{V}_{T,i,i^\star}| L(|\mathcal{V}_{T,i,i^\star}|, \delta)} \\
&= \sum_{k \in \mathcal{V}_{T,i,i^\star}} (h(x_{k,a_k^\star}) - h(x_{k,a_{k,i}})) - 0.72\sqrt{|\mathcal{V}_{T,i,i^\star}| L(|\mathcal{V}_{T,i,i^\star}|, \delta)} \\
&= R(\mathcal{V}_{T,i,i^\star}) - 0.72\sqrt{|\mathcal{V}_{T,i,i^\star}| L(|\mathcal{V}_{T,i,i^\star}|, \delta)} \,,
\end{aligned}$$

where $a_k^\star$ is the optimal prediction in round k. Combining the previous two displays allows us to bound the regret from above for any $\epsilon \in (0, 1/2)$ as

$$\begin{aligned}
R(\mathcal{V}_{T,i,i^\star}) &\leq \sum_{k \in \mathcal{V}_{T,i,i^\star}} (1 \wedge B_{k,i}) + 3\sqrt{|\mathcal{V}_{T,i,i^\star}| L(T, \delta)} + 1 \\
&\leq \sum_{k \in \mathcal{V}_{T,i,i^\star}} (1 \wedge I_{k,i} B_{k,i}) \, \mathbb{1}\{B_{k,i} \geq \epsilon\} + \frac{5}{2}\epsilon |\mathcal{V}_{T,i,i^\star}| + \frac{3}{2}\frac{L(T,\delta)}{\epsilon} + 1 \\
&\leq \frac{1}{\epsilon} \sum_{k \in \mathcal{V}_{T,i,i^\star}} (\epsilon \wedge I_{k,i} B_{k,i}^2) + \frac{5}{2}\epsilon |\mathcal{V}_{T,i,i^\star}| + \frac{3}{2}\frac{L(T,\delta)}{\epsilon} + 1 \\
&\leq 8\frac{d_i}{\epsilon} + \frac{5}{2}\epsilon |\mathcal{V}_{T,i,i^\star}| + \frac{3}{2}\frac{L(T,\delta)}{\epsilon} + 1 \,,
\end{aligned}$$

where the last inequality applies the condition of the d_i test. Since $\mathcal{V}_{T,i,i^\star}$ can only contain rounds where i was chosen and requested a label, we can apply the label complexity bound from Eq. (15) (with $\sum_{k \in [T]} p_{k,i}$ therein upper bounded as explained just afterwards) which gives

$$|\mathcal{V}_{T,i,i^\star}| = O\left(\inf_{\epsilon \in (0,1/2)} \left(\epsilon^\gamma T \left(\frac{d(S_{T,n}(h), \delta)}{d_{i_2}}\right)^{\gamma+1} + \frac{d_{i_2}}{\epsilon^2}\right) + L(T, \delta/\log T)\right) \,, \tag{19}$$

and plugging this back into the previous bound yields, for any $i = \langle i_1, i_2 \rangle$,

$$R(\mathcal{V}_{T,i,i^\star}) = O\left(\frac{d_{i_2}}{\epsilon} + T\left(\frac{\epsilon \, d(S_{T,n}(h), \delta)}{d_{i_2}}\right)^{\gamma+1} + \frac{L(T,\delta)}{\epsilon} + L(T, \delta/\log T)\right) \,.$$

Summing over $i \neq i^*$ gives the claimed result. □

A.2.3 Putting it all together

Putting together the above results gives rise to the following guarantee on the regret and the label complexity of Algorithm 2, presented in the main paper as Theorem 2.

Theorem 4. *Let Algorithm 2 be run with parameters δ, $\gamma \leq \alpha$ with a pool of base learners $\mathcal{M}_1$ of size M on an i.i.d. sample $(x_1, y_1), \dots, (x_T, y_T) \sim \mathcal{D}$, where the marginal distribution $\mathcal{D}_\mathcal{X}$ fulfills the low-noise condition with exponent $\alpha \geq 0$ w.r.t. a function h that satisfies (1) and having complexity $S_{T,n}(h)$. Let also $\mathcal{M}_1$ contain at least one base learner i such that $\sqrt{2}S_{T,n}(h) \leq S_i \leq 2\sqrt{2}S_{T,n}(h)$ and $d_i = \Theta(L_H(L_H + \log(M/\delta) + S^2_{T,n}(h)))$, where $L_H = \log\det(I + H)$, being H the NTK matrix of depth n over the set of points $\{x_{t,a}\}_{t=1,\dots,T,\, a=\pm 1}$. Also assume $m \geq CT^4 \log(2Tn/\delta) n^6 \left(T^2 \vee 1/\lambda_0^4\right)$ where C is the constant in Lemma 1 and Lemma 2. Then with probability at least $1-\delta$ the cumulative regret R_T and the total number of queries N_T are simultaneously upper bounded as follows:*

$$R_T = O\left(M\left(L_H\big(L_H + \log(M/\delta) + S^2_{T,n}(h)\big)\right)^{\gamma+1} T^{\frac{1}{\gamma+2}} + M\,L(T,\delta)\right)$$
$$N_T = O\left(M\left(L_H\big(L_H + \log(M/\delta) + S^2_{T,n}(h)\big)\right)^{\frac{\gamma}{\gamma+2}} T^{\frac{2}{\gamma+2}} + M\,L(T,\delta)\right)\,,$$

where $L(T,\delta)$ is the logarithmic term defined at the beginning of Algorithm 2's pseudocode.

Proof. Using the decomposition in Eq. (16) combined with Lemmas 13, 14, and 15 we see that the regret of Algorithm 2 can be bounded as

$$\begin{aligned}
R(T) \leq\ & R(\mathcal{R}_{i_\star}) + R(\mathcal{U}_{i_\star}) + R(\mathcal{O}_{i_\star}) \\
&= O\Bigg(\frac{M}{\epsilon} 2^{\gamma+1} d(S_{i_1^\star}, \delta)^{\gamma+2} + \frac{M}{\epsilon} 2^{\gamma+1} d(S_{i_1^\star}, \delta)^{\gamma+1} L(T,\delta) + \epsilon T_\epsilon \\
&\quad + \sum_{i=\langle i_1, i_2\rangle \in \mathcal{M}_1} \inf_{\epsilon \in (0,1/2)} \left(\frac{d_{i_2}}{\epsilon} + T\left(\frac{\epsilon\, d(S_{T,n}(h),\delta)}{d_{i_2}}\right)^{\gamma+1} + \frac{L(T,\delta)}{\epsilon}\right) + ML(T, \delta/\log T)\Bigg)\,.
\end{aligned}$$

We first bound term T_ϵ through Lemma 23 (Appendix A.4). This gives, with probability at least $1-\delta$,

$$T_\epsilon = O\left(T\epsilon^\gamma + \log\frac{\log T}{\delta}\right)\,,$$

simultaneously over ϵ. Plugging back into the above, collecting terms and resorting to a big-oh notation that disregards multiplicative constants independent of T, M, $1/\delta$ yields

$$R(T) = O\Bigg(\frac{M}{\epsilon}\Big(d(S_{T,n}(h),\delta)^{\gamma+2} + d(S_{T,n}(h),\delta)^{\gamma+1} L(T,\delta)\Big) + \epsilon^{\gamma+1} T + ML(T, \delta/\log T) \tag{20}$$

$$+ \sum_{i=\langle i_1, i_2\rangle \in \mathcal{M}_1} \inf_{\epsilon \in (0,1/2)} \left(\frac{d_{i_2}}{\epsilon} + T\left(\frac{\epsilon\, d(S_{T,n}(h),\delta)}{d_{i_2}}\right)^{\gamma+1} + \frac{L(T,\delta)}{\epsilon}\right)\Bigg)\,, \tag{21}$$

holding simultaneously for all $\epsilon \in (0, 1/2)$.

Now, the sum of the first two terms in the RHS (that is, Eq. (20)) is minimized by selecting ϵ of the form

$$\epsilon = \left(M\left(\frac{d(S_{T,n}(h),\delta)^{\gamma+2} + d(S_{T,n}(h),\delta)^{\gamma+1} L(T,\delta)}{T}\right)\right)^{\frac{1}{\gamma+2}}$$

which, plugged back into (20) gives

$$\begin{aligned}
(20) &= O\left(\left(M\left(d(S_{T,n}(h),\delta)^{\gamma+2} + d(S_{T,n}(h),\delta)^{\gamma+1} L(T,\delta)\right)\right)^{\frac{\gamma+1}{\gamma+2}} T^{\frac{1}{\gamma+2}} + ML(T,\delta/\log T)\right) \\
&= O\left(Md(S_{T,n}(h),\delta)^{\gamma+1}\, T^{\frac{1}{\gamma+2}}\, L(T,\delta/\log T)\right)\,.
\end{aligned}$$

Notice that ϵ is constrained to lie in $(0, 1/2)$. If that is not the case with the above choice of ϵ, our bound delivers vacuous regret guarantees.

As for the sum in (21), each term in the sum is individually minimized by an ϵ of the form

$$\epsilon = \left(\frac{(d_{i_2} + L(T,\delta)) \cdot d_{i_2}^{\gamma+1}}{T \cdot d(S_{T,n}(h),\delta)^{\gamma+1}} \right)^{\frac{1}{\gamma+2}}.$$

Notice that the above value of ϵ lies in the range $(0, \frac{1}{2})$ provided $d_{i_2} = o(T^{\frac{1}{\gamma+2}})$. Hence we simply assume that our model selection algorithm is performed over base learners with d_{i_2} bounded as above. In fact, if $d(S_{T,n}(h),\delta)$ exceeds this range then our bounds become vacuous.

Next, substituting the value of ϵ obtained above we get that Eq. (21) can be bounded as

$$(21) = O\left(M d(S_{T,n}(h),\delta)^{\frac{\gamma+1}{\gamma+2}} T^{\frac{1}{\gamma+2}} \right).$$

Combining the bounds on Eq. (20) and Eq. (21) we get the claimed bound on the regret R_T.

Next, we bound the label complexity of the our model selection procedure. From Lemma 12 we have that the label complexity can be bounded by

$$N_T = O\left(\sum_{i=\langle i_1,i_2\rangle \in \mathcal{M}_1} \left(\frac{d_{i_2}}{\epsilon^2} + \epsilon^\gamma T \left(1 \wedge \frac{d(S_{T,n}(h),\delta)}{d_{i_2}} \right)^{\gamma+1} \right) + ML(T,\delta/\log T) \right). \quad (22)$$

Next consider a term in the summation in Eq. (22) with $d_{i_2} \geq d(S_{T,n}(h),\delta)$. The following value of ϵ minimizes the term:

$$\epsilon = \left(\frac{d_{i_2}}{T^{\frac{1}{\gamma+2}}} d(S_{T,n}(h),\delta)^{-\frac{\gamma+1}{\gamma+2}} \right).$$

Again we notice that this is a valid range of ϵ provided that $d_{i_2} = o(T^{\frac{1}{\gamma+2}})$. Substituting back into Eq. (22) we obtain that the label complexity incurred due to such terms (denoted by $N_1(T)$) is bounded as

$$\begin{aligned} N_1(T) &= O\left(M \frac{T^{\frac{2}{\gamma+2}} d(S_{T,n}(h),\delta)^{\frac{2(\gamma+1)}{\gamma+2}}}{d_{i_2}} + ML(T,\delta/\log T) \right) \\ &= O\left(M T^{\frac{2}{\gamma+2}} d(S_{T,n}(h),\delta)^{\frac{\gamma}{\gamma+2}} + ML(T,\delta/\log T) \right). \end{aligned} \quad (23)$$

Finally, consider a term in the summation in Eq. (22) with $d_{i_2} < d(S_{T,n}(h),\delta)$. Then the value of ϵ that minimizes the term equals

$$\epsilon = \left(\frac{d_{i_2}}{T} \right)^{\frac{1}{\gamma+2}}.$$

Substituting back into Eq. (22), we get that the label complexity incurred by such terms (denoted by $N_2(T)$) is bounded by

$$N_2(T) = O\left(M T^{\frac{2}{\gamma+2}} d(S_{T,n}(h),\delta)^{\frac{\gamma}{\gamma+2}} + ML(T,\delta/\log T) \right). \quad (24)$$

Noting that $N_T = N_1(T) + N_2(T)$, we get the claimed bound on the label complexity of the algorithm. □

A.3 Extension to non-Frozen NTK

Following [45], in order to avoid computing $f(x,\theta_0)$ for each input x, we replace each vector $x_{t,a} \in \mathbb{R}^{2d}$ by $[x_{t,a}, x_{t,a}]/\sqrt{2} \in \mathbb{R}^{4d}$, matrix W_l by $\begin{pmatrix} W_l & 0 \\ 0 & W_l \end{pmatrix} \in \mathbb{R}^{4d \times 4d}$, for $l = 1, \ldots, n-1$, and W_n by $\left(W_n^\top, -W_n^\top\right)^\top \in \mathbb{R}^{2d}$. This ensures that the initial output of neural network $f(x,\theta_0)$ is always 0 for any x.

Algorithm 3: NTK Selective Sampler.

Input: Confidence level δ, complexity parameter S, network width m and depth n, number of rounds T, step size η, number of gradient descent steps J .

Initialization:

- Generate each entry of W_k independently from $\mathcal{N}(0, 4/m)$, for $k \in [n-1]$, and each entry of W_n independently from $\mathcal{N}(0, 2/m)$;
- Define $\phi_t(x) = g(x; \theta_{t-1})/\sqrt{m}$, where $\theta_{t-1} = \langle W_1, \ldots, W_n \rangle \in \mathbb{R}^p$ is the weight vector of the neural network so generated at round $t-1$;
- Set $Z_0 = I \in \mathbb{R}^{p \times p}$.

for $t = 1, 2, \ldots, T$

Observe instance $x_t \in \mathcal{X}$ and build $x_{t,a} \in \mathcal{X}^2$, for $a \in \mathcal{Y} = \{-1, +1\}$

Set $\mathcal{C}_{t-1} = \{\theta : \|\theta - \theta_{t-1}\|_{Z_{t-1}} \leq \frac{\gamma_{t-1}}{\sqrt{m}}\}$, with $\gamma_{t-1} = 3(\sqrt{\log\det Z_{t-1} + 3\log(1/\delta)} + S)$

Set

$$U_{t,a} = f(x_{t,a}, \theta_{t-1}) + \gamma_{t-1}\|\phi_{t-1}(x_{t,a})\|_{Z_{t-1}^{-1}} + \tfrac{1}{\sqrt{T}}$$

Predict $a_t = \arg\max_{a \in \mathcal{Y}} U_{t,a}$

Set $I_t = \mathbb{1}\{|U_{t,a_t} - 1/2| \leq B_t\} \in \{0, 1\}$ with $B_t = 2\gamma_{t-1}\|\phi_{t-1}(x_{t,a_t})\|_{Z_{t-1}^{-1}} + \frac{2}{\sqrt{T}}$

if $I_t = 1$

Query $y_t \in \mathcal{Y}$, and set loss $\ell_t = \ell(a_t, y_t)$

Update

$$Z_t = Z_{t-1} + \phi_t(x_{t,a_t})\phi_t(x_{t,a_t})^\top$$

$$\theta_t = \mathrm{TrainNN}\Big(\eta,\, J,\, m,\, \{x_{s,a_s} \,|\, s \in [t], I_s = 1\},\, \{\ell_s \,|\, s \in [t], I_s = 1\},\, \theta_0\Big)$$

else $Z_t = Z_{t-1},\ \theta_t = \theta_{t-1},\ \gamma_t = \gamma_{t-1},\ \mathcal{C}_t = \mathcal{C}_{t-1}$.

Algorithm 4: TrainNN(η, J, m, $\{x_i\}_{i=1}^l$, $\{\ell_i\}_{i=1}^l$, $\theta^{(0)}$)

Input: Step size η, number of gradient descent steps J, network width m, contexts $\{x_i\}_{i=1}^l$, loss values $\{\ell_i\}_{i=1}^l$, initial weight $\theta^{(0)}$.

Set $\mathcal{L}(\theta) = \sum_{i=1}^l (f(x_i, \theta) - 1 + \ell_i)^2/2 + m\|\theta - \theta^{(0)}\|_2^2$.

for $j = 0, \ldots, J-1$

$\theta^{(j+1)} = \theta^{(j)} - \eta\nabla\mathcal{L}(\theta^{(j)})$

Return $\theta^{(J)}$

A.3.1 Non-Frozen NTK Base Learner

The pseudocode for the base learner in the non-frozen case is contained in Algorithm 3. Unlike Algorithm 1, Algorithm 3 updates θ_t using gradient descent. The update of θ_t is handled by the pseudocode in Algorithm 4.

Note that both Algorithm 1 and Algorithm 3 determine the confidence ellipsoid $\mathcal{C}_t$ by updating θ_t, γ_t and Z_t. To tell apart the two learners, we use $\bar{\gamma}_t$, $\bar{Z}_t$ and $\bar{\theta}_t$ to denote the ellipsoid parameters for Algorithm 1. We make use of a few relevant lemmas from [45] and its references therein stating that in the over-parametrized regime, i.e., when $m = \mathrm{poly}(T, n, \lambda_0^{-1}, S^{-1}, \log(1/\delta))$, the gradient descent update does not leave θ_t and Z_t too far from the corresponding $\bar{\theta}_t$ and $\bar{Z}_t$. Moreover, the neural network f is close to its first order approximation. The interested reader is referred to Lemmas B.2 through B.6 of [45]. Combining these results with the analysis in Section A.1 we bound the label complexity and regret for Algorithm 3.

The below proofs are mainly sketched, since they follow from a combination of the arguments in Section A.1 and some technical lemmas in [45].

We re-define here $\mathcal{E}_0$ to be the event where (4) and (5) hold along with all the bounds in the well-approximation lemmas of [45] (Lemmas B.2 throug B.6). From [45], there exists a constant C such that if

$$m \geq CT^{19}n^{27}(\log m)^3$$

then $\mathbb{P}(\mathcal{E}_0) \geq 1-\delta$. Event $\mathcal{E}$ is defined as in Eq. (7) with this specific event $\mathcal{E}_0$ therein.

We give a new version of Lemma 3 below, which implies that event $\mathcal{E}$ still holds with high probability for Algorithm 3, with a specific learning rate η, number of gradient descent steps J and network width m.

Lemma 16. *There exist positive constants $\bar{C}_1, \bar{C}_2$ such that if*

$$\eta = \frac{\bar{C}_1}{2mnT}\,, \qquad J = \frac{4nT}{\bar{C}_1}\log\frac{S}{CnT^{3/2}}\,, \qquad m \geq \bar{C}_2 T^{19}n^{27}(\log m)^3$$

and $\sqrt{2}S_{T,n}(h) \leq S$, then under event $\mathcal{E}_0$ for any $\delta \in (0,1)$ we have with probability at least $1-\delta$

$$\|\theta^* - \theta_t\|_{Z_t} \leq \gamma_t/\sqrt{m}$$

simultaneously for all $t > 0$. In other words, under event $\mathcal{E}_0$, $\theta^* \in \mathcal{C}_t$ with high probability for all t.

Proof sketch. In Lemma 5.2 of [45], it is shown that

$$\begin{aligned}\sqrt{m}\|\theta^* - \theta_t\|_{Z_t} \leq{}& \sqrt{1 + Cm^{-1/6}\sqrt{\log m}n^4t^{7/6}} \\ &\times \left(\sqrt{\log\det Z_t + Cm^{-1/6}\sqrt{\log m}n^4t^{5/3} + 2\log(1/\delta)} + S\right) \\ &+ Cn\left((1-\eta m)^{J/2}t^{3/2} + Cm^{-1/6}\sqrt{\log m}n^{7/2}t^{19/6}\right)\end{aligned}$$

for some constant C under event $\mathcal{E}_0$ and the assumption that $\sqrt{2}S_{T,n}(h) \leq S$. Setting $\eta = \frac{\bar{C}_1}{2mnT}$ and $J = \frac{4nT}{\bar{C}_1}\log\frac{S}{CnT^{3/2}}$ allows us to bound $Cn(1-\eta m)^{J/2}T^{3/2}$ by S. Lastly, since m satisfies

$$\frac{C^2\sqrt{\log m}\,n^{9/2}T^{19/6}}{m^{1/6}} \leq 1\,,$$

we have

$$\begin{aligned}\sqrt{m}\|\theta^* - \theta_t\|_{Z_t} &\leq \sqrt{2}\left(\sqrt{\log\det Z_t + 1 + 2\log(1/\delta)} + S\right) + S + 1 \\ &\leq 3\left(\sqrt{\log\det Z_t + 3\log(1/\delta)} + S\right)\,,\end{aligned}$$

as claimed. □

We next show the properties of $\widehat{\Delta}_t$ and Δ_t, which is a new version of Lemma 6 for the non-frozen case.

Lemma 17. *Assume $m \geq poly(T, n, \lambda_0^{-1}, S, \log(1/\delta))$ and $\sqrt{2}S_{T,n}(h) \leq S$. Then under event $\mathcal{E}$ we have $0 \leq \widehat{\Delta}_t - \Delta_t \leq B_t$ and $0 \leq \widehat{\Delta}_t$, where B_t is the querying threshold in Algorithm 3, i.e.,*

$$B_t = 2\gamma_{t-1}\|\phi_t(x_{t,a_t})\|_{Z_{t-1}^{-1}} + \frac{2}{\sqrt{T}}\,.$$

Proof. Denote

$$\tilde{U}_{t,a} = \max_{\theta\in\mathcal{C}_{t-1}}\langle g(x_{t,a};\theta_{t-1}), \theta - \theta_0\rangle = \langle g(x_{t,a};\theta_{t-1}), \theta_{t-1} - \theta_0\rangle + \gamma_{t-1}\|\phi_t(x_{t,a})\|_{Z_{t-1}^{-1}}\,.$$

We decompose

$$\widehat{\Delta}_t - \Delta_t = (U_{t,a} - \tilde{U}_{t,a}) + (\tilde{U}_{t,a} - h(x_{t,a})) =: A_1 + A_2\,.$$

For A_1, by definition of $U_{t,a}$ in Algorithm 3 we have

$$U_{t,a} - \tilde{U}_{t,a} = f(x_{t,a};\theta_{t-1}) - \langle g(x_{t,a};\theta_{t-1}), \theta_{t-1} - \theta_0\rangle + \frac{1}{\sqrt{T}}\,.$$

Under event $\mathcal{E}$, the bound in Lemma B.4 of [45] holds. That is, there is a constant C_2 such that

$$\begin{aligned}|f(x_{t,a};\theta_{t-1}) - \langle g(x_{t,a};\theta_{t-1}),&\theta_{t-1}-\theta_0\rangle| \\ &= |f(x_{t,a};\theta_{t-1}) - f(x_{t,a};\theta_0) - \langle g(x_{t,a};\theta_{t-1}),\theta_{t-1}-\theta_0\rangle| \\ &\le C_2 m^{-1/6}\sqrt{\log m} n^3 t^{2/3} .\end{aligned}$$

Setting m so large as to satisfy $C_2 m^{-1/6}\sqrt{\log m} n^3 T^{2/3} \le \frac{1}{2\sqrt{T}}$ gives us

$$\frac{1}{2\sqrt{T}} \le A_1 \le \frac{3}{2\sqrt{T}} .$$

To estimate A_2 we decompose it further as

$$\begin{aligned}A_2 &= \left(\tilde{U}_{t,a} - \langle g(x_{t,a};\theta_{t-1}),\theta^\star-\theta_0\rangle\right) + \left(\langle g(x_{t,a};\theta_{t-1}),\theta^\star-\theta_0\rangle - \langle g(x_{t,a};\theta_0),\theta^\star-\theta_0\rangle\right) \\ &=: A_3 + A_4 .\end{aligned}$$

Following the argument in Lemma 6 we can show the inequality $0 \le A_3 \le 2\gamma_{t-1}\|\phi_t(x_{t,a_t})\|_{Z_{t-1}^{-1}}$ under event $\mathcal{E}$. By Cauchy-Schwartz inequality $|A_4| \le \|g(x_{t,a};\theta_{t-1}) - g(x_{t,a};\theta_0)\|_2 \|\theta^\star - \theta_0\|_2$. Using the assumption that the bounds in Lemmas B.5 and B.6 in [45] hold and $\sqrt{2}S_{T,n}(h) \le S$, there exists a constant C_1 such that

$$|A_4| \le \|g(x_{t,a};\theta_{t-1}) - g(x_{t,a};\theta_0)\|_2 \|\theta^\star-\theta_0\|_2 \le C_1 S m^{-1/6}\sqrt{\log m} n^{7/2} t^{1/6} .$$

Setting m large enough to satisfy $C_1 S m^{-1/6}\sqrt{\log m} n^{7/2} T^{1/6} \le \frac{1}{2\sqrt{T}}$ gives us

$$-\frac{1}{2\sqrt{T}} \le A_2 \le 2\gamma_{t-1}\|\phi_t(x_{t,a_t})\|_{Z_{t-1}^{-1}} + \frac{1}{2\sqrt{T}} .$$

Combining the bound for A_1 and A_2 we obtain

$$0 \le \widehat{\Delta}_t - \Delta_t \le B_t ,$$

which proves the first part of the claim.

Next, since $U_{t,a} - h(x_{t,a}) \ge 0$ for $a \in \mathcal{Y}$, we also have

$$U_{t,1} + U_{t,-1} \ge h(x_{t,1}) + h(x_{t,-1}) = 1$$

which, by definition of a_t, gives $U_{t,a_t} \ge \frac{1}{2}$, i.e., $\widehat{\Delta}_t \ge 0$. This concludes the proof. □

As a consequence of the above lemma, like in the frozen case, on rounds where Algorithm 3 does not issue a query, we are confident that prediction a_t suffers no regret.

Before bounding the label complexity and regret, we give the following lemma which is the non-frozen counterpart to Lemma 5 in Section A.1. The proof follows from very similar arguments, and is therefore omitted.

Lemma 18. *Let η, J and m be as in Lemma 16 and $\sqrt{2}S_{T,n}(h) \le S$. Then for any $b > 0$ we have*

$$\sum_{t=1}^{T} b \wedge I_t B_t^2 = O\left(\left(\log\det Z_T + \log(1/\delta) + S^2 + b\right)\log\det Z_T\right) . \tag{25}$$

Combining the above lemmas we can bound the label complexity and regret similar to Section A.1.

Lemma 19. *Let η, J be as in Lemma 16, $m \ge poly(T, n, \lambda_0^{-1}, S, \log(1/\delta))$, and $\sqrt{2}S_{T,n}(h) \le S$. Then under event $\mathcal{E}$ for any $\epsilon \in (0, 1/2)$ we have*

$$\begin{aligned}N_T &= O\left(T_\epsilon + \frac{1}{\epsilon^2}(\log\det Z_T + \log(1/\delta) + S^2)\log\det Z_T\right) \\ &= O\left(T_\epsilon + \frac{1}{\epsilon^2}\left(\log\det(I+H) + \log(1/\delta) + S^2\right)\log\det(I+H)\right) .\end{aligned}$$

Lemma 20. *Let η, J be as in Lemma 16, $m \geq poly(T, n, \lambda_0^{-1}, S, \log(1/\delta))$, and $\sqrt{2}S_{T,n}(h) \leq S$. Then under event $\mathcal{E}$ for any $\epsilon \in (0, 1/2)$ we have,*

$$R_T = O\left(\epsilon T_\epsilon + \frac{1}{\epsilon}\left(\log\det Z_T + \log(1/\delta) + S^2\right)\log\det Z_T\right)$$
$$= O\left(\epsilon T_\epsilon + \frac{1}{\epsilon}\left(\log\det(I+H) + \log(1/\delta) + S^2\right)\log\det(I+H)\right) .$$

The rest of the analysis follows from the same argument that relies on Lemma 23 (Appendix A.4) allowing one to replace T_ϵ by $O\left(T\epsilon^\alpha + O\left(\log\frac{\log T}{\delta}\right)\right)$, and culminating into a statement very similar to Theorem 1.

A.3.2 Model Selection for Non-Frozen NTK Base Learners

The pseudocode for the model selection algorithm applied to the case where the base learners are of the form of Algorithm 3 instead of Algorithm 1 is very similar to Algorithm 2, and so is the corresponding analysis. The adaptation to non-frozen base learners simply requires to change a constant. Specifically, we replace '8' in the d_i test of Algorithm 2 with '432', all the rest remains the same, provided the definition of $B_{t,i}$ (querying threshold of the i-th base learner) is now taken from Algorithm 3 (B_t therein).

An analysis very similar to Lemma 11 shows that a well-specified learner is (with high probability) not removed from the pool $\mathcal{M}_t$, while the label complexity and the regret analyses mimic the corresponding analyses contained in Section A.2.1 and A.2.2, with inflated constants and network width m.

A.4 Ancillary technical lemmas

Lemma 21. *Let $i, j \in \mathcal{M}_1$ be two base learners. with probability at least $1 - 2\delta$ the following concentration bound holds for all rounds t*

$$\left|\sum_{k\in\mathcal{V}_{t,i,j}} (\mathbb{1}\{a_{k,i} \neq y_k\} - \mathbb{1}\{a_{k,j} \neq y_k\} + h(x_{k,a_{k,i}}) - h(x_{k,a_{k,j}}))\right| \leq 0.72\sqrt{|\mathcal{V}_{t,i,j}|L(|\mathcal{V}_{t,i,j}|,\delta)} .$$

Proof. We write the LHS of the inequality to show as $\left|\sum_{k=1}^t Y_k\right|$ where

$$Y_k = \mathbb{1}\{k \in \mathcal{V}_{t,i,j}\}(\mathbb{1}\{a_{k,j} = y_k\} - \mathbb{1}\{a_{k,i} = y_k\} + h(x_{k,a_{k,i}}) - h(x_{k,a_{k,j}})).$$

and let $\mathbb{E}_k$ and Var_k denote expectation and variance conditioned on everything before y_k (including $x_k, a_{k,i}, a_{k,j}$ and i_k). Note that Y_k is a martingale difference sequence since $\mathbb{E}_k Y_k = 0$. Further, $H_k = \mathbb{1}\{k \in \mathcal{V}_{t,i,j}\}(1 + h(x_{k,a_{k,i}}) - h(x_{k,a_{k,j}}))$ and $G_k = -\mathbb{1}\{k \in \mathcal{V}_{t,i,j}\}(-1 + h(x_{k,a_{k,i}}) - h(x_{k,a_{k,j}}))$ are predictable sequences with $-G_k \leq Y_k \leq H_k$. Thus, we can apply Lemma 27 and get that with probability at least $1-\delta$, for all $t \in \mathbb{N}$

$$\sum_{i=1}^t Y_i \leq 1.44\sqrt{(W_t \vee m)\left(1.4\log\log\left(2\left(\frac{W_t}{m}\vee 1\right)\right) + \log\frac{5.2}{\delta}\right)}$$
$$\leq 0.72\sqrt{|\mathcal{V}_{t,i,j}|\left(1.4\log\log\left(2|\mathcal{V}_{t,i,j}|\right) + \log\frac{5.2}{\delta}\right)} = 0.72\sqrt{|\mathcal{V}_{t,i,j}|L(|\mathcal{V}_{t,i,j}|,\delta)}$$

where $W_t = |\mathcal{V}_{t,i,j}|/4$ and $m = 1/4$. We can apply the same argument to $-Y_k$ which yields the statement to show. □

Lemma 22. *For any $i \in \mathcal{M}_1$ the number of rounds in which i was played is bounded with probability at least $1 - \delta$ for all $t \in [T]$ as*

$$|\mathcal{T}_{t,i}| \leq \frac{3}{2}\sum_{k=1}^t p_{k,i} + 1.45L(t,\delta) .$$

Proof. We can write the size of $T_{t,i}$ by its definition as $|\mathcal{T}_{t,i}| = \sum_{k=1}^{t} \mathbb{1}\{i_k = i\}$. We denote by $\mathcal{F}_k$ the σ-field induced by all observed quantities in Algorithm 2 before i_k is sampled (including the set of active learners $\mathcal{M}_k$). By construction $(\mathcal{F}_t)_{t\in\mathbb{N}}$ is a filtration. Note further that $\mathbb{1}\{i_k = i\}$ conditioned on $\mathcal{F}_k$ is Bernoulli random variable with probability $p_{k,i}$. We can therefore apply Lemma 26 with $Y_k = \mathbb{1}\{i_k = i\} - p_{k,i}$, $m = p_{1,i}$ (which is a fixed quantity) and $W_t = \sum_{k=1}^{t} p_{k,i}(1 - p_{k,i}) \leq \sum_{k=1}^{t} p_{k,i}$. This gives that with probability at least $1 - \delta$

$$\begin{aligned}\sum_{k=1}^{t} \mathbb{1}\{i_k = i\} - \sum_{k=1}^{t} p_{k,i} \leq& 1.44\sqrt{L(t,\delta)\sum_{k=1}^{t} p_{k,i}} + 0.41L(t,\delta)\\ \leq& \frac{1}{2}\sum_{k=1}^{t} p_{k,i} + 1.45L(t,\delta).\end{aligned}$$

Note that $W_t/p_{1,i} \leq t$ holds because the smallest non-zero probability $p_{k,i}$ is $p_{1,i}$. Rearranging terms yields the desired statement. □

Lemma 23. *Under the low-noise assumption with exponent $\alpha \geq 0$, each of the following three bounds holds for any $i \in [M]$ with probability at least $1 - \log_2(12T)\delta$:*

$$\forall t \in [T], \epsilon \in (0, 1/2): \quad |\mathcal{T}^{\epsilon}_{t,i}| \leq 3\epsilon^{\alpha} \sum_{k=1}^{t} p_{k,i} + 2L(t,\delta), \tag{26}$$

$$\forall t \in [T], \epsilon \in (0, 1/2): \quad |\mathcal{T}^{\epsilon}_{t,i}| \leq 3\epsilon^{\alpha} |\mathcal{T}_{t,i}| + 2L(|\mathcal{T}_{t,i}|,\delta), \tag{27}$$

$$\epsilon \in (0, 1/2): \quad T_{\epsilon} \leq 3\epsilon^{\alpha} T + 2L(T,\delta)\ . \tag{28}$$

Proof. We here show the result for Eq. (26). The arguments for Eq. (27) and Eq. (28) follow analogously (by considering $\mathbb{1}\{i_k = i\}$ and 1 instead of $p_{k,i}$). To show Eq. (26), we first prove this condition for a *fixed* $\epsilon \in (0, 1/2]$: We begin by writing $T^{\epsilon}_{t,i}$ by its definition as

$$|\mathcal{T}^{\epsilon}_{t,i}| = \sum_{k=1}^{t} \mathbb{1}\{i_k = i\}\, \mathbb{1}\{|\Delta_k| \leq \epsilon\}\ .$$

We denote by $\mathcal{F}_k$ the σ-field induced by all quantities determined up to the end of round $k-1$ in Algorithm 2 (including the set of active learners $\mathcal{M}_k$ but not i_k or x_k). By construction $(\mathcal{F}_t)_{t\in\mathbb{N}}$ is a filtration. Conditioned on $\mathcal{F}_k$, the r.v. $\mathbb{1}\{i_k = i\}\, \mathbb{1}\{|\Delta_k| \leq \epsilon\}$ is a Bernoulli random variables with probability $q_k \leq p_{k,i}\epsilon^{\alpha}$, because the choice of learner and the distribution of $|\Delta_k| \leq \epsilon$ are independent in each round and by low noise condition, the latter is at most ϵ^{α}. We can therefore apply Lemma 26 with $Y_k = \mathbb{1}\{i_k = i\}\, \mathbb{1}\{|\Delta_k| \leq \epsilon\} - q_k$, $m = q_1$ and $W_t = \sum_{k=1}^{t} q_k(1-q_k) \leq \sum_{k=1}^{t} q_k$. This gives that with probability at least $1 - \delta$

$$\begin{aligned}\sum_{k=1}^{t} \mathbb{1}\{i_k = i\}\, \mathbb{1}\{|\Delta_k| \leq \epsilon\} - \sum_{k=1}^{t} q_k \leq& 1.44\sqrt{L(t,\delta)\sum_{k=1}^{t} q_k} + 0.41L(t,\delta)\\ \leq& \frac{1}{2}\sum_{k=1}^{t} q_k + 1.45L(t,\delta),\end{aligned}$$

where the second inequality follows from AM-GM. Rearranging terms and using $q_k \leq p_{k,i}\epsilon^{\alpha} \leq p_{k,i}$ gives for a fixed ϵ

$$|\mathcal{T}^{\epsilon}_{t,i}| \leq \frac{3}{2}\epsilon^{\alpha} \sum_{k=1}^{t} p_{k,i} + 1.45L(t,\delta)\ . \tag{29}$$

We now consider the following set of values for ϵ

$$\mathcal{K} = \left\{ \left(\frac{1}{3T}\right)^{1/\alpha} 2^{\frac{i-1}{\alpha}} : i = 1, \ldots, \left\lfloor \log_2\left(\frac{3T}{2^{\alpha-1}}\right) \right\rfloor \right\} \cup \{1/2\}\ .$$

and apply the argument above for all $\epsilon \in \mathcal{K}$ which gives that with probability at least $1 - \delta|\mathcal{K}| \geq 1 - \log_2(12T)\delta$, the bound in Eq. (29) holds for all $\epsilon \in \mathcal{K}$ and $t \in \mathbb{N}$ simultaneously. In this event, consider any arbitrary $\epsilon \in (0, 1/2)$ and $t \in [T]$. Then

$$|\mathcal{T}_{t,i}^{\epsilon}| \leq |\mathcal{T}_{t,i}^{\epsilon'}| \leq \frac{3}{2}\epsilon'^{\alpha} \sum_{k=1}^{t} p_{k,i} + 1.45L(t,\delta),$$

where $\epsilon' = \min\{x \in \mathcal{K} \colon x \geq \epsilon\}$. If ϵ' is the smallest value in $\mathcal{K}$, then $\frac{3}{2}\epsilon'^{\alpha} \sum_{k=1}^{t} p_{k,i} \leq 1/2 \leq {}^1\!/_2 L(t,\delta)$. Thus, the RHS is bounded as $2L(t,\delta)$ in this case. If ϵ' is not the smallest value in $\mathcal{K}$, then by construction $2\epsilon^{\alpha} \geq \epsilon'^{\alpha}$ and the RHS is bounded as $\frac{3}{2}\epsilon'^{\alpha} \sum_{k=1}^{t} p_{k,i} + 1.45L(t,\delta) \leq 3\epsilon^{\alpha} \sum_{k=1}^{t} p_{k,i} + 1.45L(t,\delta)$. Combining both cases gives the desired result for Eq. (26). □

Lemma 24 (Elliptical potential, Lemma C.2 [35])**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and $V_t = V_0 + \sum_{i=1}^{t} x_i x_i^{\top}$ and $b > 0$ then*

$$\sum_{t=1}^{n} b \wedge \|x_t\|^2_{V_{t-1}^{-1}} \leq \frac{b}{\log(b+1)} \log \frac{\det V_n}{\det V_0} \leq (1+b) \log \frac{\det V_n}{\det V_0}.$$

Lemma 25 (Randomized elliptical potential)**.** *Let $x_1, x_2, \cdots \in \mathbb{R}^d$ and $I_1, I_2, \cdots \in \{0,1\}$ and $V_0 \in \mathbb{R}^{d \times d}$ be random variables so that $\mathbb{E}[I_k | x_1, I_1, \ldots, x_{k-1}, I_{k-1}, x_k, V_0] = p_k$ for all $k \in \mathbb{N}$. Further, let $V_t = V_0 + \sum_{i=1}^{t} I_i x_i x_i^{\top}$. Then*

$$\sum_{t=1}^{n} b \wedge \|x_t\|^2_{V_{t-1}^{-1}} \leq 1 \vee 2.9 \frac{b}{p} \left(1.4 \log\log(2bn \vee 2) + \log \frac{5.2}{\delta}\right) + \frac{2}{p}(1+b) \log \frac{\det V_n}{\det V_0}$$

holds with probability at least $1 - \delta$ for all n simultaneously where $p = \min_k p_k$ is the smallest probability.

Proof. This proof is a slight generalization of the Lemma C.4 in [35]. We provide the full proof here for convenience: We decompose the sum of squares as

$$\sum_{t=1}^{n} b \wedge \|x_t\|^2_{V_{t-1}^{-1}} \leq \frac{1}{p} \sum_{t=1}^{n} (bI_t \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) + \sum_{t=1}^{n} \frac{1}{p_t}(p_t - I_t)(b \wedge \|x_t\|^2_{V_{t-1}^{-1}}) \tag{30}$$

The first term can be controlled using the standard elliptical potential lemma in Lemma 24 as

$$\frac{1}{p} \sum_{t=1}^{n} (bI_t \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) \leq \frac{1}{p}(1+b) \ln \frac{\det V_n}{\det V_0}.$$

For the second term, we apply an empirical variance uniform concentration bound. Let $\mathcal{F}_{i-1} = \sigma(V_0, x_1, p_1, I_1, \ldots, x_{i-1}, I_{i-1}, x_i, p_i)$ be the sigma-field up to before the i-th indicator. Let $Y_i = \frac{1}{p_i}(p_i - I_i)\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)$ which is a martingale difference sequence because $\mathbb{E}[Y_i | \mathcal{F}_{i-1}] = 0$ and consider the process $S_t = \sum_{i=1}^{t} Y_i$ with variance process

$$\begin{aligned} W_t &= \sum_{i=1}^{t} \mathbb{E}[Y_i^2 | \mathcal{F}_{i-1}] = \sum_{i=1}^{t} \frac{1}{p_i^2} \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \mathbb{E}[(p - I_i)^2 | \mathcal{F}_{i-1}] \\ &= \sum_{i=1}^{t} \frac{1 - p_i}{p_i} \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \leq \sum_{i=1}^{t} \frac{b}{p_i} \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \leq \sum_{i=1}^{t} \frac{b^2}{p_i}. \end{aligned}$$

Note that $Y_t \leq b$ and therefore, S_t satisfies with variance process W_t the sub-ψ_P condition of [22] with constant $c = b$ (see Bennett case in Table 3 of [22]). By Lemma 26 below, the bound

$$\begin{aligned} S_t \leq 1.44 \sqrt{(W_t \vee m)\left(1.4 \ln\ln(2(W_t/m \vee 1)) + \ln \frac{5.2}{\delta}\right)} \\ + 0.41b \left(1.4 \ln\ln(2(W_t/m \vee 1)) + \ln \frac{5.2}{\delta}\right) \end{aligned}$$

holds for all $t \in \mathbb{N}$ with probability at least $1-\delta$. We set $m = \frac{b}{p}$ and upper-bound the RHS further as

$$\begin{aligned}
&1.44\sqrt{\frac{b}{p}\left(1 \vee \sum_{i=1}^{t}\left(b \wedge \|x_i\|^2_{V_{i-1}^{-1}}\right)\right)\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right)} \\
&+ 0.41b\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right) \\
&\leq \frac{1}{2}\left(1 \vee \sum_{i=1}^{t}\left(b \wedge \|x_i\|^2_{V_{i-1}^{-1}}\right)\right) + 1.45\frac{b}{p}\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right),
\end{aligned}$$

where the inequality is an application of the AM-GM inequality. Thus, we have shown that with probability at least $1-\delta$, for all n, the second term in Eq. (30) is bounded as

$$\frac{1}{p}\sum_{t=1}^{n}(p_t - I_t)(b \wedge \|x_t\|^2_{V_{t-1}^{-1}}) \leq \frac{1}{2}\left(1 \vee \sum_{i=1}^{n}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)\right) + Z.$$

where $Z = 1.45\frac{b}{p}\left(1.4\ln\ln\left(2bn \vee 2\right) + \ln\frac{5.2}{\delta}\right)$. And when combining all bounds on the sum of squares term in Eq. (30), we get that either $\sum_{i=1}^{n}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \leq 1$ or

$$\begin{aligned}
\sum_{i=1}^{n}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) &\leq 2Z + \frac{2}{p}(1+b)\ln\frac{\det V_n}{\det V_0} \\
&\leq \frac{4}{p}(1+b)\ln\frac{\ln(2bn \vee 2)5.2\det V_n}{\delta \det V_0}
\end{aligned}$$

which gives the desired statement. □

Lemma 26 (Time-uniform Bernstein bound)**.** *In the terminology of [22], let $S_t = \sum_{i=1}^{t} Y_i$ be a sub-ψ_P process with parameter $c > 0$ and variance process W_t. Then with probability at least $1-\delta$ for all $t \in \mathbb{N}$*

$$\begin{aligned}
S_t \leq 1.44\sqrt{(W_t \vee m)\left(1.4\log\log\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \log\frac{5.2}{\delta}\right)} \\
+ 0.41c\left(1.4\log\log\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \log\frac{5.2}{\delta}\right)
\end{aligned}$$

where $m > 0$ is arbitrary but fixed. This holds in particular when $W_t = \sum_{i=1}^{t}\mathbb{E}_{i-1}Y^2$ and $Y_i \leq c$ for all $i \in \mathbb{N}$.

Proof. The proof follows directly from Theorem 1 with the condition in Table 3 and their stitching boundary in Eq. (10) of [22]. □

Lemma 27 (Time-uniform Hoeffding bound)**.** *Let Y_t be a a martingale difference sequence and G_t, H_t two predictable sequences such that $-G_t \leq Y_t \leq H_t$. Then with probability at least $1-\delta$ for all $t \in \mathbb{N}$*

$$\sum_{i=1}^{t} Y_i \leq 1.44\sqrt{(W_t \vee m)\left(1.4\log\log\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \log\frac{5.2}{\delta}\right)}$$

where $m > 0$ is arbitrary but fixed and $W_t = \frac{1}{4}\sum_{i=1}^{t}(G_i + H_i)^2$.

Proof. We use the results of [22]. In their terminology, Table 3 in that work shows that $\sum_{i=1}^{t} Y_i$ is a sub-ψ_N process with variance process W_t. We can thus apply their Theorem 1 with the stitching boundary in their Eq. (10) with $c = 0$. Setting $\eta = 2$ and $s = 1.4$ gives the desired result. □