Dynamic Ensemble Active Learning: A Non-Stationary Bandit with Expert Advice

Citation for published version:

Digital Object Identifier (DOI):
10.1109/ICPR.2018.8545422

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and/or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Dynamic Ensemble Active Learning: A Non-Stationary Bandit with Expert Advice

Kunkun Pang∗, Mingzhi Dong†, Yang Wu‡, Timothy M. Hospedales∗
∗University of Edinburgh, Email: k.pang@ed.ac.uk, t.hospedales@ed.ac.uk
†University College London, Email: mingzhi.dong.13@ucl.ac.uk
‡Nara Institute of Science and Technology, Email: yangwu@rsc.naist.jp

Abstract—Active learning aims to reduce annotation cost by predicting which samples are useful for a human teacher to label. However, it has become clear that there is no best active learning algorithm. Inspired by various philosophies about what constitutes a good criteria, different algorithms perform well on different datasets. This has motivated research into ensembles of active learners that learn what constitutes a good criteria in a given scenario, typically through multi-armed bandit algorithms. Though algorithm ensembles can lead to better results, they overlook the fact that not only does algorithm efficacy vary across datasets, but also, during a single active learning session. That is, the best algorithm is non-stationary. This breaks existing algorithms’ guarantees and hampers their performance. In this paper, we propose dynamic ensemble active learning as a more general and promising research direction. We develop a dynamic ensemble active learner based on a non-stationary multi-armed bandit with expert advice algorithm. Our dynamic ensemble selects the right criteria at each step of active learning. It has theoretical guarantees, and shows encouraging results on 13 popular datasets.

I. INTRODUCTION

The key barrier to scaling or applying supervised learning in practice is often the cost of obtaining sufficient annotation. Active Learning (AL) aims to address this by designing query algorithms that effectively predict which points will be useful to annotate, thus enabling efficient allocation of human annotation effort. There are many different AL algorithms, each with appealing – yet completely different – motivations for what constitutes a good question to ask underpinning their design. For example, uncertainty or margin-based sampling [1], [2] suggests querying the most uncertain or ambiguous point that is the closest point to the decision boundary. Expected error reduction [3], [4] queries points that the current model predicts will reduce its future error. Another typical approach is to label the most representative samples [5], [6], [7] to ensure the major clusters within the dataset are correctly estimated. Besides these approaches, query-by-committee active learning queries points based on the disagreement between a committee of classifiers [8], [9], [10]. More recent studies investigated hybrid criteria that balance multiple motivations [11], [12], [13].

These are all good ideas, yet there are situations where each is ineffective. For example, if the classes are heavily overlapped in an area of feature-space, uncertainty sampling will be tied up querying points in an impossible to solve region. If the current model is very poor, expected error reduction cannot accurately estimate its own future error. If the main data clusters are already well classified, representative sampling focused approaches may not fine-tune between them.

These thought experiments are reflected empirically. The best algorithm for pool-based AL in practice varies both across datasets and also with the progress of learning within a given dataset [14], [15]. This observation has motivated research into both learning dataset and time-specific weightings for an AL algorithm ensemble. [16], [17] developed heuristics for switching AL algorithms that are typically good at early versus late stage learning. In contrast, [14], [15] developed methods for rapid online learning of a dataset-specific weighting for algorithms within an AL-ensemble.

The key insight of the Combination of Active Learning Online (COMB) [14] and Active Learning by Learning (ALBL) [15] algorithms is to formalise the query criteria selection task as a multi-armed bandit (MAB) problem. MAB problems have been well studied and many powerful algorithms with optimality guarantees exist. For example, if each query criterion in the ensemble is considered to be a bandit arm, and the learning improvement after executing a criterion is considered to be the arm’s reward, then MAB algorithms such as EXP3 (Exponential-weight algorithm for Exploration and Exploitation) [18] can be applied to quickly learn the efficacy of the arms (AL criteria) and is guaranteed to achieve a near optimal overall reward (learning improvement). A variant of this is to consider data-points to be arms, and AL criteria to be experts providing advice about promising arms. Then MAB with expert advice algorithms such as EXP4.P (Exponential-weight algorithm for Exploration and Exploitation using Expert advice with high probability regret bound) [19] optimise exploration and exploitation of experts, and achieve provably near optimal reward.

The fundamental limitation of existing MAB-based approaches to AL is that their underlying MAB algorithms do not take into account the temporal dynamics of active learning: different criteria are effective at different learning stages [16], [17]. This issue is illustrated by Fig. 1(a,c,e), where the most effective criterion varies across the entire time horizon. On fourclass, Density (DE) sampling is slightly better at first and uncertainty (US) is consistently good later.

This work was supported by JSPS KAKENHI Grant Number 15K16024. It has been accepted at ICPR 2018 and won Piero Zamperoni Best Student Paper Award. The corresponding author is Yang Wu.
on. Similarly in ILPD or \textit{german}, representative (RS) and density (DE) sampling are better at the crucial early stage before uncertainty becomes better. A second issue is that the scale of an accuracy-based reward falls dramatically over time (Fig. 1(d,f)). Because of this stationary bandit learners will be unduly biased by the high reward from an initial observation and fail to adapt subsequently. For example in ILDP a stationary learner may fail to make the switch from DE to US as later rewards in favour of US are small in scale compared to the initial reward in favour of DE.

Therefore there are non-stationary aspects both in reward scale, and in reward distribution per-arm (MAB perspective) or per-expert (MAB with expert advice perspective). Thus the MAB problem is formally non-stationary, violating a fundamental assumption required to guarantee existing MAB algorithms’ optimality bounds.

Here we develop a performance guaranteed stochastic MAB with expert advice\footnote{We use terminology from \cite{18}. It also has other names, including ‘contextual bandit’ \cite{19}, \cite{20}, ‘partial-label problem’ \cite{21}, and ‘associative bandit problem’ \cite{22}.} algorithm in a non-stationary environment. Applying this to AL means that, like \cite{15}, if there is a single best (but a priori unknown) AL algorithm for a dataset, we are able to quickly discover it and thus approach the performance of an oracle that knows the best algorithm for each dataset. But importantly when different algorithms’ efficacies vary over time within one dataset, we can adapt to this and approach the performance of an oracle that knows the best AL algorithm \textit{at each iteration}.

\section{Background and Related Work}

\subsection{Active Learning}

We denote the pool of data with $M$ samples as $\mathcal{D} = \{x_1, x_2, \ldots, x_M\}$ where the instances are $x_i \in \mathbb{R}^d$ and the labels are $l_i \in \{1, \ldots, C\}$. In an active learning scenario, the data $\mathcal{D}$ are initially a labelled set $\mathcal{L}$ and unlabelled set $\mathcal{U} = \mathcal{D} \setminus \mathcal{L}$ where $|\mathcal{L}| \ll |\mathcal{U}|$. Training an initial classifier $f_0$ on the samples in the initial set $\mathcal{L}$, the algorithm starts to query instances $x_i$ from $\mathcal{U}$ during iterations $t = 1, \ldots, T$. After the supervision $l_q$ of instance $x_q$ is obtained, $x_q$ is removed from the unlabelled set $\mathcal{U}$ and added to the labelled set $\mathcal{L}$, from which classifier $f_t$ is retrained.

\subsection{Bandit Algorithms}

\textbf{Multi-armed Bandit} In multi-armed bandit (MAB) problems, a player pulls a lever from a set $\mathcal{K} = \{1, \ldots, K\}$ of slot machines in a sequence of time steps $T = \{1, \ldots, T\}$ to maximise her payoff. During the game, she only observes the reward $r^k(t) \in [0, 1]$ of the specific arm pulled $k$ at time step $t$. The aim of the player is to maximise their return, which is the sum of the rewards over the sequence of pulls. This requires a trade-off between exploration (collect information to estimate the arm with the highest return) and exploitation (focus on the arm with the highest estimated return). Training a bandit learner to solve a MAB problem is then formalized as minimising the \textit{regret} between the actions chosen by the player’s strategy $\alpha_k \sim \pi$, and the best arm.

For example, the EXP3 algorithm \cite{18} minimises, for any finite $T$, the “\textit{static regret}” between the player’s reward and the best arm in retrospect: $\max_k \sum_{t=1}^{T} r^k(t) - \mathbb{E}(\sum_{t=1}^{T} r^\pi(t))$.

\textbf{Contextual Multi-armed Bandit} The goal of contextual bandits is to build a relationship between available context information $h_t \in \mathbb{R}^d$ and the reward distribution over all arms. For example, LinUCB \cite{23} makes the linear realizability assumption that there exists an unknown weight vector $\theta^* \in \mathbb{R}^d$ with $||\theta^*|| \leq 1$ so that regret $\sum_{t=1}^{T} r_{a_t}(t) - \sum_{t=1}^{T} r_{a_t}(t)$ is minimized, where $r_{a_t}(t) = \theta^\top h_t$ and $r_{a_t}(t) = \theta^\top h_t$. However, learning to predict the reward for each data point accurately appears to be an even harder problem given the
limited information from only expert suggestions (Fig 1). More importantly, given the changing reward distribution over time, there is no constant relation between context and reward.

**Multi-armed Bandit with Expert Advice** Expert information about the likely efficacy of each arm is often available. [18] thus introduced an adversarial MAB with expert advice algorithm EXP4 that exploits $N$ experts giving advice vectors (probabilities $\xi^n(t) \in [0,1]^K$ over levers) to the learner at each time. In contrast to MAB without expert advice, the goal is now to identify the best expert rather than the best arm. In this setting the regret to minimise is the difference between the return of the best expert in retrospect and the player:

$$\max_n \sum_{t=1}^T y^n_t - \mathbb{E}\left(\sum_{t=1}^T y^*_t\right)$$

where $y^n_t = \sum_{k=1}^K \xi^n_k(t) \times r^k(t)$ is the expected reward of an expert and $y^*_t$ is the expected reward of our policy.

**C. Bandits for Active Learning**

For active learning using a MAB with expert advice algorithm, the $N$ experts correspond to our ensemble of active learning criteria and the $K$ arms are available points in the pool. Each expert (criterion) $n$ provides a probability vector encoding preference $\xi^n(t)$ over arms (instances). Active learners based on MAB with expert advice aim to learn the best criterion for a specific dataset. In COMB [14], the authors propose to use MAB with expert advice in active learning and heuristically designed the classification entropy maximization (CEM) score as the reward of the EXP4 bandit algorithm [18]. A more recent paper [15] (ALBL) proposed to replace the CEM reward with an unbiased estimation of test accuracy Important Weighted Accuracy (IWA) and used an upgraded bandit algorithm EXP4.P [19], which improves the earlier EXP4 method. Similarly, another recent paper [24] applied linear upper confidence bound contextual bandit algorithm (LinUCB) to train an ensemble and transferred the knowledge to other datasets. All of these algorithms enable the selection of a suitable active learning criteria for a given dataset. Our contribution is also to perform AL in a dataset-specific way by optimally tuning the exploration and exploitation of an ensemble of AL algorithms; but more importantly to do so dynamically, thus allowing the optimal tuning to vary as learning progresses. Unlike [14], [15], [24] we are able to deal with the non-stationary nature of this process. And unlike the heuristics in [16], [17], we have a theoretical guarantees, and can work with more than two criteria.

**D. Non-stationary Property of Active Learning**

**Demonstration of Non-stationarity** We describe a preliminary experiment to demonstrate empirically the existence of non-stationary reward distributions for a MAB formalisation of AL. Following the learning trajectory of our method, we use an oracle to score all the available query points at each iteration (i.e., hypothetically label each point, update the classifier, and check the test accuracy). Using the actual test accuracy as the reward, we can obtain the true expected reward of the $n$th expert $y^n_t = \xi^n(t) r(t)$ at each time step $t$. Fig. 1 summarises the resulting average reward obtained in every 10 iterations of AL. Based on this, we can further compute the proportion of times that each criterion would obtain the highest reward. It can be seen that the MAB problem is non-stationary as the rewards vary systematically, and there is not a single criterion (expert) which obtains the highest proportion of wins throughout learning. Additionally, the ideal combination of criteria varies across datasets. For example, as illustrated in Fig 1, density and uncertainty sampling show better complementary in ILPD, while representative and uncertainty sampling are more complementary in german dataset.

**Existing MAB ensembles are not robust to non-stationarity**

The non-stationary property in the MAB formalisation of AL also highlights the key weakness of COMB and ALBL: they use EXP4/EXP4.P [18], [19] expert advice bandit algorithms which provide guarantees against an inappropriate (static) regret that is only relevant in a stationary problem. In a non-stationary problem, it is clear that even an algorithm that perfectly estimates the best single expert (optimal w.r.t static oracle Eq. 1) can be arbitrarily worse than one which can choose the best expert at each step (optimal w.r.t dynamic oracle).

In this paper, we develop a non-stationary stochastic MAB algorithm REXP4 (Restarting Exponential-weight algorithm for Exploration and Exploitation using Expert advice) with bounds against a stricter dynamic oracle notion of optimality more suited for (non-stationary) AL.

**Prior attempts at non-stationary active learners** A few previous active learning studies also observed that different algorithms are effective at different stages of learning and proposed heuristics for switching two base query criteria (e.g., density sampling at an early stage, and uncertainty sampling later on) [16], [17]. But these only adapt 2 criteria (density and uncertainty) unlike MAB ensembles which learn to combine many criteria, and their heuristics do not provide a principled and optimal way to learn when to switch.

**Prior attempts at non-stationary MABs** Some previous studies have extended MAB without expert advice learning to the non-stationary setting [25], [26] and provided regret bounds to guarantee the algorithms’ performance. However bandits with expert advice are preferable because they can achieve tighter learning bounds [18], [19] and they do not treat each criterion as a black box, so that one observation can be informative about many arms. Consider an AL situation where two criteria prefer the same instance. In the MAB interpretation (criteria=arms), after observing a reward, you only learn about the criterion/arm chosen at that iteration. In the MAB with expert advice interpretation (criteria=experts), the observed reward generates updates about the efficacy of all criteria that expressed opinions about the point.

Those few MABs extended to the non-stationary setting have other stronger assumptions. For example, the discounted/sliding-window UCB algorithm [25] assumes the
nature of the non-stationarity is that the reward distribution is piece-wise and the number of changes is known. Similarly [27] makes the easier piecewise assumption, and also that the retrospective rewards for un-pulled arms are available – but they are not in active learning. In [28], the authors proposed to measure the total statistical variance of the consecutive distributions at each time interval. Their result provides a big picture of the regret landscape for full information and bandit settings. Their proposed method addresses non-stationary environments but only for the regular MAB problem. Despite the use of the term expert in the title, it does not address the Expert-advice variant of the MAB problem relevant to us. It addresses arms rather than experts over arms.

We propose a non-stationary MAB with expert advice algorithm that has performance guarantees, and validate its practical application to active learning.

III. NON-STATIONARY MULTI-ARMED BANDIT WITH EXPERT ADVICE FOR ACTIVE LEARNING

A. Non-stationary Multi-Armed Bandit with Expert Advice: REXP4

To formalise the problem, we assume the expected reward $y^n_t$ of each expert $n$ can change at any time step $t$. The total variation of the expected reward over all $T$ steps is

$$\sum_{t=1}^{T-1} \sup_n |y^n_t - y^n_{t+1}|$$ (2)

Following [29], [26], we assume this total variation in expected reward is bounded by a variation budget $V_T$. The variation budget captures our assumed constraints on the non-stationary environment. It allows a wide variety of reward changes – from continuous drift to discrete jumps – yet provides sufficient constraint to permit a bandit algorithm to learn in a non-stationary environment. Temporal uncertainty set $V$ is defined as the set of reward vector sequences that are subject to the variation budget $V_T$ over all $T$ steps.

$$V = \left\{ y \in [0,1]^{N \times K} : \sum_{t=1}^{T-1} \sup_n |y^n_t - y^n_{t+1}| \leq V_T \right\}$$

To bound the performance of a bandit learner in a non-stationary environment, we work with the regret between the learner and a dynamic oracle. The regret is defined as the worst-case difference between the expected policy return and the return of using the best expert at each time $t$.

Definition 1. Dynamic Regret for Multi-Armed Bandit with Expert Advice

$$R^x(V,T) = \sup_{y \in V} \left\{ \sum_{t=1}^{T} y^n_t - \mathbb{E}_{\pi}^{\pi} \left[ \sum_{t=1}^{T} y^n_t \right] \right\}$$ (3)

where $y^n_t = \max_n y^n_t$ is the best possible expected reward among all experts at time $t$. Our regret is against this dynamic oracle, in contrast to prior MABs’ static oracle (Eq 1).

Our non-stationary MAB with expert advice algorithm REXP4 minimises the dynamic regret in Eq 3. As shown in Algorithm 1, it trades off between the need to remember and forget by breaking the task into batches and applying EXP4 [18] on each batch. As the reward distribution changes, it adapts to the change as by re-estimating each expert’s reward distribution at each batch. We show the worst case bound on the regret between this REXP4 procedure and the dynamic oracle.

B. Regret Bound for REXP4

The regret bound for REXP4 is illustrated in the following theorem. The theorem is proved by following the proof structure of [26] and replacing the term $\mu$ in [26] with the expected reward term $y$ in our paper.

Theorem 1. Let $\pi$ be the REXP4 policy with a batch size $\Delta_T = \lceil (A \log N)^{1/3} (T/V_T)^{2/3} \rceil$ and $\gamma = \min \{1, \sqrt{\frac{A \log N}{(c+1)\Delta_T}}\}$. Then, there is some constant $C$ such that for every $T \geq 1$, $K \geq 2$, $N \geq 2$, and $V_T \in [A^{-1}, A^{-1}T]$

$$R^x(V,T) \leq C (A \log N \cdot V_T)^{1/3} T^{2/3}$$ (4)

where $A = \min \{N,K\}$ indicates the smaller number of experts or arms.

The result is an upper bound on the regret between our REXP4 policy and the dynamic oracle. As $A = \min \{N,K\}$, it is favourable if either the number of experts $N$ or arms $K$ is small. This also means it is relatively robust to many arms (as in AL, where arms=data points). If $V_T$ is sub-linear in $T$ (total variation in reward grows slower than timesteps), then performance converges to that of the oracle.

C. Dynamic Ensemble Active Learning

Based on our REXP4 algorithm for MAB with expert advice, we present DEAL-REXP4 (Dynamic Ensemble Active Learning) for active learning based on REXP4. Our dynamic ensemble learner will update both base learner $f_t$ and active criteria weights $w(t)$ iteratively. More specifically, each ensemble criterion will predict scores $s^n_t$ for all unlabelled instances. We use exponential ranking normalisation $\exp(-\alpha \text{ rank})$ to avoid the issue of different criterion scales, and apply the Gibbs measure $\sum_n \exp(-\beta s^n_t)$ where the parameters $\alpha, \beta$ control the sharpness of the distribution. The $\text{rank}$ denotes the ranking position of the instance’s score where the ranking order is determined by the criterion strategy’s ordering. For example, the entropy criterion prefers
Algorithm 2 DEAL: Dynamic Ensemble Active Learning

Inputs: $\gamma \in (0, 1]$, initial weight $w(1) = 1$, $\Delta_T = 10$, $\tau = 1$; labelled set $\mathcal{L}_0$, unlabelled set $\mathcal{U}_0$, initial classifier $f_0$. for $t = 1 \rightarrow T$ do
1) Get scores of instance $s^n_i$ from criteria
2) Normalise the score vector $s^t_i = \frac{\exp(-\alpha \text{rank})}{\sum_k \exp(-\beta s^t_i)}$
3) Obtain the advise vector with $\xi^n_t(t) = \sum_k \exp(-\beta s^t_i)\gamma$
4) Set $W_t = \sum_{n=1}^{N} w_n(t)$ and for $k = 1, \ldots, K$ set
   $p_k(t) = (1 - \gamma) \sum_{n=1}^{N} \frac{w_n(t)\xi^n_k(t)}{W_t} + \gamma$
5) Query the label of instance $x_{k_i}$ randomly from $\mathcal{U}_t$ according to probability $p_1(t), \ldots, p_K(t)$
6) Move the instance $x_{k_i}$ from $\mathcal{U}_t$ to $\mathcal{L}_t$
7) Retrain the classifier $f_t$ and receive reward $r^k_t \in [0, 1]
8) For $k = 1, \ldots, K$ set
   $\hat{r}_k(t) = \begin{cases} r_k(t)/p_k(t) & \text{if } k = k_t \\ 0 & \text{otherwise} \end{cases}$
9) For $n = 1, \ldots, N$ set
   $\hat{y}^n_i = \xi^n(t)\hat{r}(t)$
   $w_n(t + 1) = w_n(t) \exp(\gamma \hat{y}^n_i / K)$
10) $\tau = \tau + 1$
     if $\tau > \Delta_T$ then Reset $\tau = 1$ and $w(t + 1) = 1$
end if
end for

points with maximum entropy, so the maximum entropy point has rank 1. Similarly, the minimum margin criterion prefers points with low distance to margin, so the minimum distance point has rank 1. Based on the current suggestions from the criteria members, the active learning ensemble $w(t)$ will select an instance for label querying. Then, the base learner $f_{t+1}$ will be updated with the new labelled data and the active learner $w(t + 1)$ will be updated successively based on the performance improvement of the updated base learner. To learn the non-stationary reward distribution, we use our proposed REXP4 algorithm to learn the weights of active learning criteria in an online adaptive way by introducing the restart scheme. Giving the current within-batch index $\tau \in \{1, \ldots, \Delta_T\}$, the restart scheme will be activated when $\tau > \Delta_T$, otherwise updates follow the EXP4 rule. The details are described in Algorithm 2 with an illustration in Fig. 2.

In DEAL-REXP4 we set the reward as the resulting accuracy after a classifier update. Thus in the context of active learning, the bound given in Eq. 4 means that we know that the total area under the reward curve obtained by DEAL-REXP4 is within a bound of the best case scenario that would occur only if we had known the best criterion to use at each iteration. Moreover, if the variation budget $V_T$ grows sub-linearly with $T$, DEAL-REXP4 converges towards this best-expert-per-iteration upper bound scenario.

D. Discussion of Static and Dynamic Active Learning

We divide active learning algorithms into static/dynamic based on the stationary/non-stationary assumption on the importance of each criterion over different time periods.

Static Active Learning Single criterion algorithms are all static, since they solve active learning with only one criterion. Regarding active learning algorithms with multiple motivations: if they are formalised as a single fixed mixture of criteria, they are also static. Since the coefficients of different motivations are fixed over all time steps, they assume that a single weighted combination is suitable at any learning stage. For example, Query Informative and Representative Examples (QUIRE) [11], Learning Active Learning (LAL) [33], and Discriminative and Representative Queries for Batch Mode Active Learning (BMDR) [12] are static active algorithms with multiple motivations.

Previously proposed ensemble algorithms ALBL [15], COMB [14], and Linear Strategy Aggregation (LSA) [24] are also static in the sense that, although the weight proportion
TABLE I: Summary of Active Learning Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Bandit Regret</th>
<th>Stationarity</th>
<th>Importance of Criterion</th>
<th>Ensemble Members</th>
<th>Property</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEAL</td>
<td>REXP4</td>
<td>Non-stationary</td>
<td>Dynamic best</td>
<td>Any Criteria</td>
<td>Dynamic</td>
</tr>
</tbody>
</table>

The ensemble of base learners includes:

US: picking the instances with max-entropy (min margin) instance in binary class datasets [1], [30] or minimum Best-versus-Second-Best (BvSB) [31] in multiclass datasets. RS: clustering the points near the margin [32] then scoring unlabelled points by their distances to the largest centroid. Distance-Furthest-First (DFF): Focuses on exploration by selecting the furthest unlabelled instance to the nearest labeled instance in binary class datasets [1], [30] or minimum Best-versus-Second-Best (BvSB) [31] in multiclass datasets. US: taking the nearest labelled point per class. Each experiment is repeated 200 times and the average testing accuracy is reported.

Criteria Ensemble: The ensemble of base learners includes: US: picking the instances with max-entropy (min margin) instance in binary class datasets [1], [30] or minimum Best-versus-Second-Best (BvSB) [31] in multiclass datasets. RS: clustering the points near the margin [32] then scoring unlabelled points by their distances to the largest centroid. Distance-Furthest-First (DFF): Focuses on exploration by selecting the furthest unlabelled instance to the nearest labeled instance in binary class datasets [35]. We use DFF which selects the furthest unlabelled instance to the nearest labelled instance [35] to replace the RS in multiclass datasets as originally RS is designed to work with binary class datasets.

...
(Theorem 1). A dataset with stationary reward distribution would increase. A dataset with stationary reward distribution would imply a corresponding variation budget for any given criterion at each batch.

The results show the number of wins/ties/losses of DEAL-REXP4 with state-of-the-art alternatives to tuning an AL-ensemble. Examples comparing the performance of DEAL and individual criteria in the ensemble are shown in Fig. 3. There is no single criterion that works best for all datasets, moreover different criteria are effective at different stages of learning. While DEAL is not best across all datasets and all time-steps (this would require the actual dynamic oracle upper bound), it performs well overall. This is summarised quantitatively across all 13 datasets in Tab. II. Each method’s performance is evaluated by the area under the learning curve at different proportions of added instances. The results show the number of wins/ties/losses of DEAL versus the alternative ensemble member of specified highest rank according to two-sided t-test. This shows for example that DEAL often ties with the top-ranked ensemble member (30 draws vs 1st rank), is usually at least as good as the second ranked member (50 wins and 45 ties vs only 35 losses) and is never the worst (0 losses vs 4th rank).

**Comparison vs State-of-the-Art** We compare our DEAL-REXP4 with state-of-the-art alternatives to tuning an AL-ensemble. Sometimes DUAL performs well, but it is highly variable depending on whether the criterion switch heuristic makes a good choice or not, as seen in Fig. 4. Tab. III summarises the results across all datasets in terms of AUC wins/draws/losses of each approach against the alternatives. DUAL has a lower row-total as it is defined for binary probability only; not for binary classifications. DUAL is engineered for a specific pair of criteria, so we apply its version using Uncertainty Sampling and Density-Weighted Uncertainty Sampling. It is also only defined for binary classification problems unlike the others.

**DEAL-REXP4 Settings:** For reward, we follow [15], [24] in using the IWA for unbiased estimation of test accuracy. To produce probabilistic preferences for points from all AL criteria, we use exponential ranking normalisation and a Gibbs measure with $\alpha = 0.1, \beta = 100$. We use batch size $\Delta_T = 10$ throughout. The choice $\Delta_T = 10$ is based on observing the typical coarse duration of performance gaps among different criteria. For example, RS wins first 20 iterations in Fig. 3(b). The reason for parameterizing in terms of $\Delta_T$ rather than $V_T$ is that it has intuitive meaning in AL context (batch-size), yet implies a corresponding variation budget for any given $T$ (Theorem 1).

**Characterising dataset (non)stationarity:** We first investigate each dataset to characterise its (non)stationarity. We use our DEAL trajectory, and use an oracle to measure the % wins of each criterion at each batch $\Delta_T$ in terms of performance increase. A dataset with stationary reward distribution would tend to have a consistent winner, and vice-versa. Although (non)stationarity is a continuum, we will describe a dataset as stationary if at least two criteria have a fraction of wins above threshold $\theta = 10\%$.

**DEAL versus Individual Criteria** Examples comparing the performance of DEAL and individual criteria in the ensemble are shown in Fig. 3. There is no single criterion that works best for all datasets, moreover different criteria are effective at different stages of learning. While DEAL is not best across all datasets and all time-steps (this would require the actual dynamic oracle upper bound), it performs well overall. This is summarised quantitatively across all 13 datasets in Tab. II. Each method’s performance is evaluated by the area under the learning curve at different proportions of added instances. The results show the number of wins/ties/losses of DEAL versus the alternative ensemble member of specified highest rank according to two-sided t-test. This shows for example that DEAL often ties with the top-ranked ensemble member (30 draws vs 1st rank), is usually at least as good as the second ranked member (50 wins and 45 ties vs only 35 losses) and is never the worst (0 losses vs 4th rank).

**Comparison vs State-of-the-Art** We compare our DEAL-REXP4 with state-of-the-art alternatives to tuning an AL-ensemble. Sometimes DUAL performs well, but it is highly variable depending on whether the criterion switch heuristic makes a good choice or not, as seen in Fig. 4. Tab. III summarises the results across all datasets in terms of AUC wins/draws/losses of each approach against the alternatives. DUAL has a lower row-total as it is defined for binary problems only, not evaluated on wine and letter datasets. The main observation is that DEAL outperforms the alternatives particularly on non-stationary datasets. On stationary datasets we are only slightly worse than ALBL. This is expected as REXP4 performs forgetting in order to adapt to changes in expert efficacy, meaning that we cannot exploit the best criterion as aggressively as ALBL’s EXP4.P MAB learner. Nevertheless, overall DEAL is fairly robust to stationary
datasets (small margin behind ALBL), while ALBL is not robust to non-stationary datasets (larger margin behind DEAL).

V. CONCLUSION

We proposed a non-stationary multi-armed bandit with expert advice algorithm REXP4, and demonstrated its application to online learning of a criterion ensemble in active learning. The theoretical results provide bounds on REXP4's optimality. The empirical results show that active learning with DEAL-REXP4 tends to perform near the best criterion in the ensemble. It performs comparable to state of the art alternative ensembles on stationary datasets, and outperforms them on non-stationary datasets.

REFERENCES


