Active Learning for Global Problems

FMJH Grant Application

James Cheshire, james.cheshire@telecom-paris.fr

October 2023

Thank you for taking the time to consider this proposal for the FMJH scholarship. This research proposal consists in developing further a promising subject I have recently initiated with Prof. Stephan Clemencon at the LTCI lab, Telecom Paris. The project would be under the theme "Mathématiques pour l'Intelligence Artificielle " and is focused on **active bipartite ranking**, with extensions to other **global learning problems**. In contrast to batch learning, where the learner is given a set of training examples in advance, in an active learning setting, training examples are given sequentially and typically, for each sample in turn, the learner has an degree of choice as to which point of the feature space the sample will be drawn. The elevator pitch for this project is to take a fundamental problem in statistical learning theory, that of bipartite ranking - for which the vast majority of related literature is solely in the batch setting, and develop a framework under the regime of active learning. Active learning has become a powerful tool, and from the classical example of clinical trials to Alpha Go, it has opened up novel applications of machine learning. Many statistical problems, traditionally considered in a batch setting are now being studied under an active regime, however, global ranking problems, such as bipartite ranking, remain relatively under looked. Bipartite ranking has a wide range of application, from design of search engines in information retrieval to medical diagnosis through credit-risk screening or anomaly detection in signal processing. Our objective is to work in the interface of mathematics, using active learning to elaborate new, more efficient, technology, while maintaining reliability through mathematical guarantees.

Before continuing with this research proposal, I will take the time to introduce myself and Prof. Stephan Clemencon. I am currently a postdoc at Telecom Paris, as part of the S2A team, under the supervision of Prof. Stephan Clemencon. I completed my PhD at the Otto-von-Guericke Universty, Magdeburg, in the summer of 2022, under the supervision of Prof. Alexandra Carpentier, with Prof. Christophe Giraud acting as second reviewer. My research interests are in the area of active learning, and during my PhD, were focused largely on multi armed bandits. I have several publications on multi armed bandits, at some of the most respected conferences in my field, COLT, ICML, Neurips - see my google scholar for details. I believe my expertise in this area makes me well suited for this project. Indeed, as we shall see in this proposal, there are strong links between multi armed bandits and active bipartite ranking, with overlap in both techniques and results. Furthermore, I am motivated to expand my knowledge in active learning, and for me, this project provides a great opportunity to both learn new techniques and apply my past experience. Prof. Stephan Clemencon is head of the S2A team, in the LTCI lab at Telecom Paris. I believe he is the ideal candidate to supervise this project, having written several seminal papers on bipartite ranking and being active in multiple areas of active learning. His links to industry - participation in the industrial chair Data Science and AI for Digitalized Industry and Services and holder of the industrial chair ML4BGD (2013-2018), also put him in an excellent position to identify directions of research that have good practical interest. We would both be very open to potential collaboration with researchers from the member institutions of the FMJH. For instance the work of Prof. Christophe Giraud, in the Probability and Statistics team, Mathematiques Orsay, has potential links to global active learning problems via active clustering. Also, Nicolas Vayatis at Department Mathématiques, ENS Paris Saclay, has considerable experience in both bipartite ranking and other global learning problems, such as change point detection.

We have already seen some success in the area of active bipartite ranking, our initial paper on the subject titled "Active Biparite Ranking", has been accepted to Neurips 2023. As mentioned, prior to this project active bipartite ranking has been largely unexplored and we believe our initial success only highlights the large potential for future research. This proposal will begin with an introduction to bipartite ranking. We will then go on to describe an initial active learning setting for bipartite

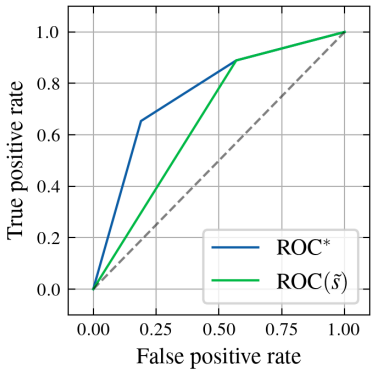

Figure 1: The ROC curves of η and $\tilde{s}$, as defined in Equations (1), with $(\beta_1, \beta_2, \beta_3) = (0.1, 0.2, 0.6)$ and C_1, C_2, C_3 equally sized.

ranking and present several of our results in this setting. The remaining majority of the proposal will then be our road map for future research.

1 Preliminaries

Bipartite Ranking The bipartite ranking problem, in the batch setting, is described as follows. The learner observes a "batch", that is, $n \geq 1$ independent copies $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of a generic random pair (X, Y) with unknown distribution, where Y is a binary random label, valued in $\{-1, +1\}$ say, and X is a high dimensional random vector, taking its values in $\mathcal{X} \subset \mathbb{R}^d$ with $d \geq 1$, that models some information hopefully useful to predict Y. The goal of the learner is then - not to assign a label, positive or negative, to any new input observation as in binary classification, X but to rank any new set of (temporarily unlabeled) observations $X'_1, \ldots, X'_{n'}$ by means of a (measurable) scoring function $s : \mathcal{X} \to \mathbb{R}$, so that those with positive label appear on top of the list (*i.e.* are those with the highest scores) with high probability.

The ROC curve criterion The ROC curve provides a popular criterion for evaluating the capacity, of a given scoring rule, to discriminate between two populations. It is used in a variety of applications, including medical diagnosis, and has received much discussion recently in diagnosis of Covid-19, see for example [33] and [34]. The ROC is the PP-plot $t \in \mathbb{R} \mapsto (1 - H_s(t),\ 1 - G_s(t))$, where $H_s(t) = \mathbb{P}\{s(X) \leq t \mid Y = -1\}$ and $G_s(t) = \mathbb{P}\{s(X) \leq t \mid Y = +1\}$, for all $t \in \mathbb{R}$. The curve can also be viewed as the graph of the càd-làg function $\alpha \in (0, 1) \mapsto \mathrm{ROC}(s, \alpha) = 1 - G_s \circ H_s^{-1}(1 - \alpha)$. The notion of ROC curve defines a partial order on the set of all scoring functions (respectively, the set of all preorders on $\mathcal{X}$): s_1 is more accurate than s_2 when $\mathrm{ROC}(s_2, \alpha) \leq \mathrm{ROC}(s_1, \alpha)$ for all $\alpha \in (0, 1)$, highlighting the global aspect of the ROC criterion. As can be proved by a straightforward Neyman-Pearson argument, the set $\mathcal{S}^*$ of optimal scoring functions is composed of increasing transforms of the posterior probability $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$. We have $\mathcal{S}^* = \{s \in \S :\ \forall (x,\ x') \in \mathcal{X}^2,\ \ \eta(x) < \eta(x') \Rightarrow s^*(x) < s^*(x')\}$ and

$$\forall (s, s^*) \in \mathcal{S} \times \mathcal{S}^*,\ \forall \alpha \in (0, 1),\ \ \mathrm{ROC}(s, \alpha) \leq \mathrm{ROC}^*(\alpha) := \mathrm{ROC}(s^*, \alpha).$$

The ranking performance of a candidate $s \in \mathcal{S}$ can be thus measured by the distance in sup-norm between its ROC curve and ROC^*, namely $d_\infty(s, s^*) := \sup_{\alpha \in (0,1)} \{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s, \alpha)\}$. For instance, consider a partition of $[0, 1]$, $\mathcal{P} = \{C_1, C_2, C_3\}$, and increasing sequence $(\beta_1, \beta_2, \beta_3) \in [0, 1]^3$ with

$$\eta(x) = \sum_{i=1}^{3} \mathbb{I}(x \in C_i)\beta_i, \qquad \tilde{s}(x) = \mathbb{I}(x \in C_1) + 2\mathbb{I}(x \in \{C_2 \cup C_3\})\ , \tag{1}$$

Where the scoring function function $\tilde{s}$ essentially, treats all points of C_2, C_3 of the same rank. See Figure 1 for the ROC curves of η and $\hat{s}$ and visualisation of the d_∞ distance. Though easy to formulate, this problem encompasses many applications, ranging from credit risk screening to the design of decision support tools for medical diagnosis through (supervised) anomaly detection. Hence, motivated by a wide variety of applications, bipartite ranking has received much attention these last few years. Many approaches to this global learning problem (*i.e.* the problem of learning a preorder on the input space $\mathcal{X}$ based on a binary feedback) have been proposed and investigated, see [11], [10], [12], [1], [9] and [31].

Active Bipartite Ranking Whereas the vast majority of dedicated articles in bipartite ranking consider the *batch* situation solely, where the learning procedure fully relies on a set $\mathcal{D}_n$ of training examples given in advance, the goal of this project is to develop an *active learning* framework for bipartite ranking, in other words to investigate this problem in an iterative context,

where the learner can formulate queries in a sequential manner, so as to observe the labels at new data points in order to refine progressively the scoring/ranking model. Precisely, the challenge consists in determining an incremental experimental design to label the data points in $\mathcal{X}$ that would permit to improve the ROC curve progressively, with statistical guarantees. Over the past years several problems in statistical learning have been investigated in a batch setting, for instance A/B testing [23], clustering [8], [28], [16], [25][35] and also change point detection [21], [20]. However, bipartite ranking remains a fundamental problem in statistical learning theory, with little consideration in the active setting.

We propose the following setting, for active bipartite ranking. The active learner plays a game with multiple time steps, where, at time each step n, they must choose a point $a_n \in \mathcal{X}$ to query, so as to observe the random label $Y_n \sim \mathcal{B}\text{er}(\eta(a_n))$ and refine the scoring model incrementally. After a sufficient number of rounds has elapsed, chosen at the learner's discretion, a final scoring function $\hat{s}$, is output. A popular criterion to evaluate the success of an active learner is to work in the fixed confidence regime, where the learner must be "probably approximately correct". That is, for some tolerance ε and probability δ, the outputted scoring function $\hat{s}$ must be such that $d_\infty(s, s^*) \leq \varepsilon$ with probability greater than $1-\delta$. Such a guarantee is termed PAC(ε, δ). Our objective is to then develop PAC(ε, δ) algorithms, with minimal expected total number of samples. While the fixed confidence regime is hugely popular in the active learning literature, there are other ways to formulate an active setting. Mainly the fixed budget setting, here the time horizon of the game is fixed, the objective of the learner is to then minimise regret with their limited budget.

2 Active bipartite ranking for piecewise constant scoring functions

Naturally, if no assumption is made on the posterior η, the problem becomes unfeasible – consider for instance the case where η alternates between 0 and 1 in arbitrarily small increments. A natural solution is to assume η is piecewise constant. That is, we consider the simplest scoring functions, measurable functions that are constant on pieces of the input space $\mathcal{X}$ forming a partition. For the entirety of this section we assume $\mathcal{X} = [0, 1)$ and introduce the grid points $\{G_1, ..., G_K\} = \{i/K : \ i = 1, \ \ldots, \ K-1\}$, where $K \geq 1$. However, our analysis holds for higher dimension, as one can just project a high dimensional grid onto the $[0, 1)$ interval. We will see that in the perspectives of future research, application to higher dimensions would not be so straight forward. Our assumption on η is then as follows,

Assumption 2.1. There exists a permutation σ of $[K]$ and distinct constants $\mu_1, \ \ldots, \ \mu_K$ in $(0, 1)$ such that

$$\eta(x) = \sum_{i=1}^{K} \mu_i \cdot \mathbb{I}\{x \in [G_{\sigma(i)}, \ G_{\sigma(i+1)})\} \text{ for all } x \in [0, 1) \ .$$

We write $p = \frac{1}{K}\sum_{i\in[K]} \mu_i$. We point out that, as η may remain constant over multiple sections of the grid, the permutation σ satisfying assumption 2.1, is not necessarily unique. For now, the parameter K is supposed to be known, in contrast with the μ_i's, which have to be learned by means of an active strategy. Such an assumption may seem restrictive, however it has precedent within the bipartite ranking literature - as shown in [15] (see subsection 2.3 therein), when smooth enough, ROC* can be accurately approximated by the (stepwise) ROC curve of a piece wise constant scoring function, and perhaps more notably, puts us in line with the multi armed bandit. Indeed our problem can now be viewed as a K armed bandit problem. Specifically, our problem is a "pure exploration bandit problem", see [6], as we have no notion of cumulative reward, and rather wish to uncover a fundamental property of the arms, or in our terminology, the posterior η. Other related pure exploration bandit problems exist, such as best arm identification, see [2], [4] and [19], and the TopM problem, where the objective of the learner is to output a list of the M best arms, see [22]. As we shall see, the global nature of the ranking problem presents several additional challenges, not typically seen in multi armed bandits.

Problem complexity The complexity of a problem is related to the expected minimum number of samples a policy must draw to be PAC(δ, ε), a quantity which depends upon the features of the problem, specifically, the shape of the posterior η. When defining our measure of problem complexity we must capture this dependence as succinctly as possible. A naive approach to the active bipartite ranking problem, is to treat each pair of points on the grid, $i, j \in [K]$ as a separate classification problem. To correctly distinguish the situations, $\mathcal{H}_0^{i,j} := \mu_i > \mu_j$, $\mathcal{H}_1^{i,j} := \mu_i < \mu_j$, with probability greater than $1-\delta$, it is well known, see e.g. [23], that for small δ, the minimum number of samples required is of the order $\frac{\log(1/\delta)}{\text{kl}^*(\mu_j, \mu_i)}$, where we remind the reader kl* is the the Chernoff Information, closely related to the kl divergence. Thus, if the learner wished to output a scoring function in S^*, the sample complexity would be of the order, $\sum_{i\in[K]} \frac{\log(1/\delta)}{\min_{j\in[K]}(\text{kl}^*(\mu_j, \mu_i))}$. Of course, distinguishing between $\mathcal{H}_0^{i,j}$ and $\mathcal{H}_1^{i,j}$ is impractical when μ_i and μ_j are very close, or even equal. However, in our regime, the learner is not required to correctly rank every pair of points $i, j \in [K]$, only to output a scoring function $\hat{s}$ such that, $d_\infty(\hat{s}, \eta) \leq \varepsilon$. Intuition indicates that the learner may be irreverent to the ranking within certain groups of points on the gird, as long as their posterior values are sufficiently close. Exactly when one is able to be ambiguous on the ranking between points relies on several factors. We show that the learner must be *at least* able to distinguish $\mathcal{H}_0^{i,j}$ vs $\mathcal{H}_1^{i,j}$, for all $j : |\mu_i - \mu_k| \leq \Delta_i$, where,

$$\Delta_i := \min\left\{ x > 0 : \sum_{i\neq j} x\mathbb{I}\big(|\mu_i - \mu_j| \leq x\big) \geq K\varepsilon p(1-\mu_i) \right\} . \tag{2}$$

For a point $i \in [K]$ we then refer to Δ_i as the gap of point i. As one can see from Equation (2), the size of Δ_i depends on several factors. Firstly there is the position of the point i relative to the ROC curve. The learner must be much more careful on areas of the feature space which correspond to the start of the ROC curve, as slight variations here can cause large regret. This affect is captured in the $p(1-\mu_i)$ term, the greater the value of point i in comparison to the overall proportion p, the tighter the necessary confidence interval around i. There is of course the effect of ε, smaller epsilon corresponding to tighter confidence intervals. The effect of K is misleading, the right hand term suggests larger K corresponds to greater gaps, however, for larger K the summation $\sum_{i \neq j} x\mathbb{I}\big(|\mu_i - \mu_j| \leq x\big)$ will typically be larger. This summation also shows a dependency on the behaviour of η in the neighbourhood of G_i. This type of dependency is not typically found in bandits, where the gaps usually depend upon the distance to a single arm, e.g. the optimal arm. With the above in mind we then define our measure of problem complexity as,

$$H_i^{(1)} = \frac{1}{\mathrm{kl}(\mu_i, \mu_i + \Delta_i) \wedge \mathrm{kl}(\mu_i, \mu_i - \Delta_i)} \, .$$

An algorithm for active bipartite ranking and upper bound on expected stopping time Once the learner has drawn one or more samples from a section of the grid, $[G_i, G_{i+1}]$, they are able to generate an empirical mean for μ_i and a confidence interval that will contain μ_i with high probability. Our algorithm `active-rank` maintains an active set of points across several rounds. At the beginning of each round `active-rank` draws a sample, uniformly, from sections of the grid remaining in the active set. The empirical mean and confidence interval at each grid point is then updated and at the end of each round, points are eliminated from the active set based on a specific criterion. We prove that `active-rank` is PAC(ε, δ) and further more demonstrate the following upper bound on it's expected stopping time.

$$c \sum_{i \in [K]} H_i^{(2)} \log\Big(c' H_i^{(2)} K^2 / \delta\Big) \, ,$$

for absolute constats c, c', where $H_i^{(2)} = \max_{j \in [K]} \Big(\frac{1}{\mathrm{kl}(\mu_j, \mu_j + \Delta_i/8)} \vee \frac{1}{\mathrm{kl}(\mu_j, \mu_j - \Delta_i/8)} \Big)$. The difference between $H_i^{(2)}$ and $H_i^{(1)}$ is due to the fact that for Bernoulli random variables, for a fixed number of samples, the width of KL divergence based confidence intervals change across the $[0, 1)$ interval. Confidence intervals closer to 0 or 1 being tighter. In cases where posterior gets very close to 0 or 1, at certain points but not others, our upper bound fails to capture this dependency. Elimination algorithms such as `active-rank` have seen wide usage in the literature for BAI, see [32],[30], [17] and [18]. However, closer to our work is the Racing algorithm [24], designed for the TopM problem, where, as in our approach, the confidence bounds used are based on the kl divergence as opposed to Hoeffdings. Their elimination criterion, however, differs considerably to our own. For simplicity let us consider the Top1 problem, that is, best arm identification - the following arguments can be extended in the case of TopM. The racing algorithm of [24] eliminates an arm $i \in [K]$ from the active set, at time t, when, the positive gap between the lower confidence bound around the highest empirical mean and the upper confidence bound at point i is greater than ε. However, due to the global nature of the ranking problem, in our setting, the decision to remove a point from the active set is not made based on the distance to another single point. We rather consider a condition on the local smoothness of the posterior around the point i. An additional difficulty that arises here is that the local smoothness around a point can potentially depend upon points no longer in the active set and once a point is no longer in the active set, we essentially have no control on the width of its confidence interval.

A lower bound on expected stopping time for active bipartite ranking We also demonstrate lower bound for the bipartite ranking problem. If p is not too small and the maximum variance of η is bounded by some constant, we demonstrate the a lower bound on the expected stopping time of any PAC(ε, δ) algorithm of the order $\sum_{i \in [K]} H_i^{(1)}$. Our bound follows from a novel application of a Fano type inequality.

3 Proposed research objectives

As bipartite ranking is a fundamental problem in statistical learning theory, and has seen little consideration in an active setting, there are many potential directions for future research, including links to other global active learning problems. Our goal in this project is to focus on settings which reflect well practical situations. We will start by discussing the limitations of our algorithm, `active-rank` and what would be needed to improve its viability in practically relevant settings. Following this we will consider several other learning objectives that expand upon the d_∞ norm and give rise to cross domain connections with clustering. Finally, we consider complete reformulations of the problem, one by replacing the piecewise constant assumption by a smoothness constraint and another by adding a structural assumption to the posterior η, each drastically changing the nature of the problem.

3.1 Tackling practical limitations of the `active-rank` algorithm

Our `active-rank` algorithm is a great first step, being optimal or near optimal in many situations. However, as we shall see, the particular class of problems it performs poorly on, may well be those that come up often in practice. Therefore it is of

great interest to replace `active-rank` with a more practically applicable, but still theoretically sound, approach. This is a multi faceted challenge, which we will now describe in detail.

Auto corrective bipartite ranking There are essentially two components in the gap between our current upper and lower bounds. The first is the additional logarithmic dependency upon K present in our upper bound. Despite being logarithmic this dependence is potentially significant, as in practical situations, the size of grid needed, for the assumption that the posterior η is piecewise constant, may be very large. If one also has sparsity on the grid - η is 0 or near 0 at almost all points, the $\log(K)$ term could then become the dominating term in our bound. The reason the $\log(K)$ term appears is due to the fact that, once `active-rank` removes a section of the grid from the active set, it will never sample it again. A similar problem is solved in [7], where the authors use a auto correcting binary search - their algorithm is able to backtrack in the binary search and correct its self, to solve a structured bandit problem. Tree based ranking procedures are well studied, see [11], and a backtracking version of such a method may be the key to removing the $\log(K)$ dependency.

Achieving true problem complexity The second component, in the gap between upper and lower bounds is, for a given $i \in [K]$, the difference in the $H_i^{(1)}$ and $H_i^{(2)}$ terms. The reason $H_i^{(2)}$ appears in our upper bound is that, the decision to remove a point $i \in [K]$ from our active set is made based on the minimum width of confidence interval across the entire grid as opposed to the local width at i. As we are dealing with Bernoulli distributions and kl divergence based confidence bounds, for a fixed number of samples, points close to zero or one will have tighter confidence bounds and thus may be sampled more than is necessary. If one were to assume that the posterior η exists solely in the interval $[\gamma, 1-\gamma]$ for some $\gamma > 0$, then for all $i \in [K]$, $H_i^{(2)}$ and $H_i^{(1)}$ will be with a constant factor of each other, with that constant depending on γ. In practical data sets one will often have huge areas of the feature space where η is near 0 or 1. It is of our opinion that it is possible to capture the true complexity in all cases, with a non-trivial modification to our proof.

3.2 The limitations of the d_∞ norm - expanding our utility

While the d_∞ provides a natural, straight forward way to measure the regret of the learner, there remain other intuitive objectives that can be useful in application.

Clustering while ranking Our algorithm `active-rank` provides a ranking of the space $[0, 1)$ by providing a permutation on $[K]$ to rank each point of the grid separately. While the PAC(δ, ε) is preserved, many pairwise rankings may be incorrect as their corresponding values in the posterior η are too close to distinguish. Therefore, while the algorithm performs well in terms of d_∞ regret, it is failing to represent an inherent structure of the posterior η. In many situations, with practical relevance, instead of ranking every single grid point in $[K]$, it will suffice to consider an ordered partition of $[0, 1]$, of cardinality less that K, i.e. for some ordered partition $\{\mathcal{P} = \{C_1, ..., C_M\}\}$, with $M \leq K$ consider the following score function,

$$\hat{s}_{\mathcal{P}}(x) = \sum_{i=1}^{M} i\mathbb{I}(x \in C_i) \ .$$

The objective will then be to output an ordered partition of minimal cardinality, in as few samples as possible, while maintaining the PAC(ε, δ) guarantee. We believe this objective is feasible but would require a none trivial adaptation of `active-rank`, one that extends beyond a sequential elimination approach and is able to eliminate large sections of the interval concurrently.

Removing the need for a fixed tolerance ε When ε becomes small enough, the ranking problem becomes equivalent to the identification of all constant sections of the posterior η. This then puts us inline with recent literature in the field of active clustering, specifically, [35] in which the authors consider a active clustering problem, represented as K armed d dimensional bandit, where the arms are split into M clusters. They work in a PAC(δ) setup where their goal is to recover the entire clustering of the arms, with probability greater than $1 - \delta$, in as few samples as possible. Comparing to our setting, if one is to view a section of the grid on which η is constant as a single cluster, by retrieving the clustering of the arms one can then easily do ranking. Their results differ to our own in several key ways though. Firstly their algorithm takes the number of clusters M as a parameter, this highlights the main difference between their setting and our own. In the Bipartite ranking problem, assuming ε is not very small, one does not have to recover exactly all the clusters to ensure regret under the d_∞ norm is less than ε. Therefore we do not need to know the number of clusters and our algorithm must be able to exploit larger ε to achieve smaller stopping times. The second key difference is that the results of [35] are hold only in the asymptotics, that is as $\delta \to 0$. Their algorithm employs a forced exploration phase, which ensures each arm is pulled at at least a sub linear rate. Essentially, this means that in such an asymptotic setting, *the means of the arms are known to the learner*, which naturally drastically changes the nature of their results. Extension to bounds for fixed $\delta > 0$ would be none trivial, noted as potential future work in [35], and essential if one were to compare to our confidence setting. With the above work in mind, an interesting objective would be to consider an algorithm, which works for "all ε" that is one which runs without a particular

stopping time in mind, but rather maintains a PAC guarantee where the tolerance ε grows optimally with sampling time. The idea being to have a unified approach, achieving optimal results both in our regime and that of [35].

Ranking the TopM Often practitioners are concerned only with ranking the top performing areas of the feature space. With similarities to the TopM problem in multi armed bandits, consider the setting where the goal of the learner is to both identify and rank the M best grid points, according to their posterior values. Naturally one would wish to solve both problems in tandem and delicate treatment would be required to correctly identify the class of problems for which ranking the top M arms becomes harder than their identification.

3.3 Removing the piecewise constant assumption

The piecewise constant assumption can be cumbersome in practice and a natural step would be to replace the assumption that the posterior η is piece wise constant on a grid of size K by a smoothness assumption on η. Such an extension would be akin to the extension of K-armed bandits to continuous armed bandits, where the arm set $[K]$ is replaced by a function, with smoothness constraint, on the $[0,1]$ interval. In continuous bandits one often makes the assumption that the function is sufficiently regular around the optimal arm. This has been well studied for cumulative regret, see [29], [3] but more related to our setting are the following works considering best arm identification, [5] and [22]. However, as we are in a global setting there is not a single point of interest, as with the optimal arm. We can instead use a global smoothness constraint, i.e. Holder continuity. For some $\beta > 0$, the posterior η is β-Holder continuous, on the interval $[0,1]$ if and only if, there exists a constant $C > 0$ such that $\forall, x, y \in [0,1]$,

$$|\eta(x) - \eta(y)| \leq |x - y|^{\beta} \ .$$

The objective of the learner would be to provide a continuous ranking of the entire interval, $\hat{\eta}$. In practice the learner would provide a ranking over a finite set of points in $[0,1]$ and rank $[0,1]$ according to its disctretisation on this finite set of points. In the bandit literature, a typical strategy is, given knowledge of β, to first discretise the $[0,1]$ interval and then apply classical techniques. In our setting, however, we conjecture such an approach would be insufficient and rather one is required to vary the level of discretisation across the $[0,1]$ interval, as due to the interconnected nature of the bipartite ranking problem, the learner will need to discretise to a higher level at certain points of the feature space. We conjecture that with the following definition of the gap of a point $y \in [0,1]$,

$$\Delta(y) := \min \left\{ x > 0 : x\lambda(\{y : \mathbb{I}\big(|\eta(x) - \eta(y)| \leq x\big) \geq \varepsilon p(1 - \eta(y))\right\} \ ,$$

were δ is the Lebesgue measure on $[0,1]$. The minimum number of samples needed by the learner would then be of the order,

$$\int_0^1 \frac{\Delta(y)^{-\beta}}{\mathrm{kl}(\eta(y) - \Delta(y), \eta(y) + \Delta(y))} \, dy \ .$$

Assuming the above conjecture holds there would remain two challenging questions in this setting. The first is how to deal with the case where β is unknown, as is common in practice. Working instead with a lower bound on β, which one then pays through a log term may be feasible. There is also the idea of estimating β, as is proposed in the setting of [5]. The second challenge that arises in this setting is extension to higher dimensions, which can no longer be done by simply projecting a d-dimensional gird onto the $[0,1)$ interval.

3.4 Structural assumptions on the posterior

In many practical applications the ranking of a specific point on the feature space will be locally dependent, e.g. spatially points are clustered with those of similar posterior value. Therefore it is of interest to consider various structural assumptions one could make on the posterior η. One such approach from the bandit literature would be spectral bandits. In this setting the arm set $[K]$ is endowed with a graph structure and the arms are assumed to be smooth on this graph - arms which are neighbours on the graph have similar values. This setting has seen recent interest for pure exploration bandit problems, see [26] and [27]. In this setting, we would consider η, not as a piece wise constant function on the grid of size K, but rather as a K sized graph, with each node corresponding to a posterior value. The posterior would then be smooth according to this graph. A key advantage of this assumption is in its generality as one can consider any structure able to be encoded into a K sized graph.

3.5 Continuous and multipartite ranking

In addition to bipartite ranking, the problem of mulitpartite ranking has been studied, see [13] and [14]. In this situation Y takes at least three ordinal values and an extension of the ROC curve criterion to the ROC manifold. A further extension would be then to consider continuous ranking. As with bipartite ranking, both these settings can be studied in a active setting.

References

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *J. Mach. Learn. Res.*, 6:393–425, 2005.

[2] J.-Y. Audibert, S. Bubeck, and R. Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer, 2010.

[3] T. Bonald and A. Proutiere. Two-target algorithms for infinite-armed bandits with bernoulli rewards. *Advances in Neural Information Processing Systems*, 26:2184–2192, 2013.

[4] A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.

[5] A. Carpentier and M. Valko. Simple regret for infinitely many armed bandits. In *International Conference on Machine Learning*, pages 1133–1141, 2015.

[6] S. Chen, T. Lin, I. King, M. R. Lyu, and W. Chen. Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems*, 27:379–387, 2014.

[7] J. Cheshire, P. Menard, and A. Carpentier. Problem dependent view on structured thresholding bandit problems. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1846–1854. PMLR, 18–24 Jul 2021.

[8] A. Choromanska and C. Monteleoni. Online clustering with experts. In *Artificial Intelligence and Statistics*, pages 227–235. PMLR, 2012.

[9] S. Clémençon and S. Robbiano. Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of ICML*, number 1, 2011.

[10] S. Clémençon and N. Vayatis. Overlaying classifiers: a practical approach for optimal scoring. *Constructive Approximation*, .:., 2008.

[11] S. Clémençon and N. Vayatis. Tree-based ranking methods. *IEEE Transactions on Information Theory*, 55(9):4316–4336, 2009.

[12] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and Empirical Minimization of U-Statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

[13] S. Clémençon and S. Robbiano. The treerank tournament algorithm for multipartite ranking. *Journal of Nonparametric Statistics*, 27(1):107–126, 2015.

[14] S. Clémençon, S. Robbiano, and N. Vayatis. Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91:67–104, 2013.

[15] S. Clemencon and N. Vayatis. On partitioning rules for bipartite ranking. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 97–104, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[16] V. Cohen-Addad, B. Guedj, V. Kanade, and G. Rom. Online k-means clustering. In *International Conference on Artificial Intelligence and Statistics*, pages 1126–1134. PMLR, 2021.

[17] E. Even-Dar, S. Mannor, and Y. Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

[18] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

[19] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[20] A. Gopalan, B. Lakshminarayanan, and V. Saligrama. Bandit quickest changepoint detection. *Advances in Neural Information Processing Systems*, 34:29064–29073, 2021.

[21] S. Hayashi, Y. Kawahara, and H. Kashima. Active change-point detection. In *Asian Conference on Machine Learning*, pages 1017–1032. PMLR, 2019.

[22] S. Kalyanakrishnan, A. Tewari, P. Auer, and P. Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

[23] E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of a/b testing. In *Conference on Learning Theory*, pages 461–481. PMLR, 2014.

[24] E. Kaufmann and S. Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pages 228–251. PMLR, 2013.

[25] A. Khaleghi, D. Ryabko, J. Mary, and P. Preux. Online clustering of processes. In *Artificial Intelligence and Statistics*, pages 601–609. PMLR, 2012.

[26] T. Kocák and A. Garivier. Epsilon best arm identification in spectral bandits. In *IJCAI*, pages 2636–2642, 2021.

[27] T. Kocák and A. Garivier. Epsilon best arm identification in spectral bandits. In Z.-H. Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2636–2642. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[28] E. Liberty, R. Sriharsha, and M. Sviridenko. An algorithm for online k-means clustering. In *2016 Proceedings of the eighteenth workshop on algorithm engineering and experiments (ALENEX)*, pages 81–89. SIAM, 2016.

[29] A. Locatelli and A. Carpentier. Adaptivity to smoothness in x-armed bandits. *31st Annual Conference on Learning Theory*, 75:1–30, 2018.

[30] S. Mannor and J. N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

[31] A. Menon and R. Williamsson. Bipartite ranking: A risk theoretic perspective. *Journal of Machine Learning Research*, 7:1–102, 2016.

[32] E. Paulson. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, 1964.

[33] C. H. Sudre, B. Murray, T. Varsavsky, M. S. Graham, R. S. Penfold, R. C. Bowyer, J. C. Pujol, K. Klaser, M. Antonelli, L. S. Canas, et al. Attributes and predictors of long covid. *Nature medicine*, 27(4):626–631, 2021.

[34] Y. Sun, V. Koh, K. Marimuthu, O. T. Ng, B. Young, S. Vasoo, M. Chan, V. J. Lee, P. P. De, T. Barkham, et al. Epidemiological and clinical predictors of covid-19. *Clinical Infectious Diseases*, 71(15):786–792, 2020.

[35] J. Yang, Z. Zhong, and V. Y. Tan. Optimal clustering with bandit feedback. *arXiv preprint arXiv:2202.04294*, 2022.