Active Bipartite Ranking with Smooth Posterior Distributions

James Cheshire
Telecom Paris

Stephan Clémençon
Telecom Paris

Abstract

In this article, bipartite ranking, a statistical learning problem involved in many applications and widely studied in the passive context, is approached in a much more general *active setting* than the discrete one previously considered in the literature. While the latter assumes that the conditional distribution is piece wise constant, the framework we develop permits in contrast to deal with continuous conditional distributions, provided that they fulfill a Hölder smoothness constraint. We first show that a naive approach based on discretisation at a uniform level, fixed *a priori* and consisting in applying next the active strategy designed for the discrete setting generally fails. Instead, we propose a novel algorithm, referred to as `smooth-rank` and designed for the continuous setting, which aims to minimise the distance between the ROC curve of the estimated ranking rule and the optimal one w.r.t. the $\sup$ norm. We show that, for a fixed confidence level $\varepsilon > 0$ and probability $\delta \in (0, 1)$, `smooth-rank` is PAC(ε, δ). In addition, we provide a problem dependent upper bound on the expected sampling time of `smooth-rank` and establish a problem dependent lower bound on the expected sampling time of any PAC(ε, δ) algorithm. Beyond the theoretical analysis carried out, numerical results are presented, providing solid empirical evidence of the performance of the algorithm proposed, which compares favorably with alternative approaches.

1 INTRODUCTION

Whether in the medical field (diagnostic assistance), in signal processing (anomaly detection), in finance (credit-risk screening) or in automatic document retrieval (search engines), the goal pursued is not always to learn how to predict a binary label Y, valued in $\{0, 1\}$ say (*e.g.* ill *vs* healthy, abnormal *vs* normal, default *vs* repayment, irrelevant *vs* relevant) based on some related random input information X like in binary classification, the flagship problem in machine-learning. More often, the aim is to learn a (real valued) ranking function $f(x)$ so as to order all possible values x for X like any increasing transform of the posterior probability $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ would do. This statistical learning problem is known as *Bipartite Ranking* and, due to the wide range of its applications, has received much attention these last few years. Because of its global (and not local) nature, the gold standard to measure ranking performance is of functional nature, namely the P-P plot referred to as the ROC curve, *i.e.* the plot of the true positive rate $\mathbb{P}(f(X) > t \mid Y = 1)$ against the false negative rate $\mathbb{P}(f(X) > t \mid Y = 0)$ as the threshold $t \in \mathbb{R}$ varies. Most of the works documented in the literature consider the passive batch learning situation, where the ranking function $f(x)$ is first built based on the preliminary observation of $n \geq 1$ independent copies of the generic random pair (X, Y) in the training stage and next applied to new (temporarily) unlabeled observations in the test/predictive phase. Various algorithmic approaches and dedicated theoretical guarantees have been elaborated, based on direct optimization of the empirical ROC curve (see Clémençon and Vayatis (2009) or Clémençon and Vayatis (2008) for instance) or on maximization of scalar summary criteria such as the AUC (*i.e.* the empirical Area Under the ROC Curve), refer to *e.g.* Agarwal et al. (2005), Clémençon et al. (2008) or Menon and Williamsson (2016). Various extensions of the original framework have been recently studied: *Multipartite Ranking*, which corresponds to the case where the ordinal label Y takes a (finite) number of values strictly larger than 2, is considered in Clémençon et al. (2013), the situation, referred to as as *Continuous Ranking*), where Y is a continuous real valued r.v. is investigated in Clémençon and Achab (2017) while the case where no label is available, known as *Unsupervised Ranking*, is investigated in Clémençon and Thomas (2018).

In bipartite ranking, the statistical framework is as follows. One observes $n \geq 1$ independent copies

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

$\{(X_1, Y_1),\ \ldots,\ (X_n, Y_n)\}$ of a generic random pair $(X, Y) \in [0,1] \times \{0,1\}$. We assume X is draw in uniformly from $[0,1]$ and models some information hopefully useful to predict Y. Specifically, $Y \sim \mathcal{D}_X$, for some unknown distribution $\mathcal{D}_X$. In contrast to classification, the goal pursued is of global (and not local) nature. It is not to assign a label, positive or negative, to any new input observation X but to rank any new set of (temporarily unlabeled) observations $X'_1,\ \ldots,\ X'_{n'}$ by means of a (measurable) scoring function $s : [0,1] \to \mathbb{R}$, so that those with greater labels appear on top of the list (*i.e.* are those with the highest scores) with high probability. More formally, the accuracy of any scoring rule can be evaluated through the ROC curve criterion or its popular scalar summary, the AUC (standing for the Area Under the ROC Curve), and, as expected, optimal scoring functions w.r.t. these performance measures can be shown to be increasing transforms of the posterior probability $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$. 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. The vast majority of dedicated articles consider the *batch* situation solely, where the learning procedure fully relies on a set of training examples given in advance, however, more recently in Cheshire et al. (2023), bipartite ranking has been considered in an *active learning* framework. In such an active setting the learner is able to 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. In the work of Cheshire et al. (2023) a key (restrictive) assumption is made that the regression function η is piece wise constant on a grid of *known* size K on the feature space. In one dimension, this means the learner assumes there are some $(\mu_k)_{k \in [K]} \in [0,1]^K$ such that:

$$\eta(x) = \sum_{i=1}^{K} \mu_k \mathbb{I}(x \in [i/K, (i+1)/K]) \,. \tag{1}$$

As pointed out in Cheshire et al. (2023), we see that under the assumption (1), the problem becomes equivalent to a *multi armed bandit*. In the multi armed bandit problem, the learner is presented with a set of K "arms", each one corresponding to some unknown distribution. The learner then plays a game of several rounds, where in each round the learner must choose one of the K arms from which they receive a sample from the corresponding distribution, see Lattimore and Szepesvári (2020) for an overview. Much of the literature on multi armed bandits concerns maximisation of cumulative reward, however, several strands of literature consider so called "pure exploration" bandit problems, where the learner does not care about there cumulative reward, but rather aims to uncover some underlying property of the arms. The classical example of a pure exploration bandit problem is best arm identification. We see that, as the size of the grid is assumed to be known to the learner, the problem under assumption (1) is then equivalent to a K armed pure exploration bandit problem - one can view each piece wise constant section of the grid as an arm. Here the goal of the learner is to uncover the ranking of the arms, with their regret given by the ROC criterion.

A fundamental problem, recognised in the bandit literature, is that, assuming one has a finite set of arms, which can all be sampled in reasonable time, is unsuitable in practice. There have been various responses to this problem, one being continuous armed bandits, otherwise known as $\mathcal{X}$-armed bandits. Here, instead of a finite set of arms, one considers a function on some feature space $\mathcal{X}$. The actions of the learner now correspond to querying points of the feature space and receiving noisy evaluations of the function at said points. Many works have considered the classical bandit problems of minimising cumulative regret and best arm identification, in the $\mathcal{X}$-armed bandit setting, see Bubeck et al. (2011), Grill et al. (2015),Bartlett et al. (2019), Akhavan et al. (2020) and Locatelli and Carpentier (2018). As the above works have considered classical bandit problems in the $\mathcal{X}$-armed setting, our goal in this paper is to consider the problem of bipartite ranking in a continuous setting. That is, we remove the piece wise constant assumption on η, which is rather assumed to be a continuous function, subject to certain smoothness constraints, as in the classic $\mathcal{X}$-armed bandit setting.

Our contributions In this paper we consider the bipartite ranking problem under a smoothness constraint on the posterior η. Specifically, we remove the piece wise constant assumption of Cheshire et al. (2023) and instead assume that the regression function η is β-Hölder smooth in each dimension, see Assumption 1, where *β is known to the learner*. We describe an algorithm, `smooth-rank`, which sequentially queries points of the feature space $\mathcal{X}$. Given a confidence level $\varepsilon > 0$ and probability $\delta > 0$, the goal of `smooth-rank` is, in a few queries as possible, to output a ranking of $\mathcal{X}$, such that the induced ROC curve of said ranking is within ε of the optimal ROC curve, in terms of the sup norm, with probability greater than $1 - \delta$. Theorem then shows that `smooth-rank` satisfies the above statistical guarantee and provides an upper bound on it's expected total number of queries. In Theorem 2, we also demonstrate a lower bound on the expected number of queries of any PAC(ε, δ) algorithm, matching the upper bound of Theorem 3.1, up to log terms.

2 BACKGROUND AND PRELIMINARIES

Notation Here we introduce several dedicated notions that will be extensively used in the subsequent analysis. By λ is meant the Lebesgue measure on $[0,1]^d$. For any $a,\ b$ in $[0,1]$, $\mathcal{B}\mathrm{er}(a)$ refers to the Bernoulli distribution with mean a and $\mathrm{kl}(a,b)$ to the Kullback Leibler divergence between the Bernoulli distributions $\mathcal{B}\mathrm{er}(a)$ and $\mathcal{B}\mathrm{er}(b)$. The indicator function of any event ξ is denoted by $\mathbb{I}\{\xi\}$, for $d \in \mathbb{N}, \varepsilon > 0$, we write $\mathcal{E}_d(\varepsilon)$ for the ε-net on $[0,1]^d$ and for $x \in [0,1]^d$, $\mathcal{B}^d_x(y)$ for the d-dimensional ball of radius y centered at x.

Bipartite ranking A rigorous formulation of ranking involves functional performance measures. Let $\mathcal{S}$ be the set of all scoring functions, any $s \in \mathcal{S}$ defines a preorder $\preceq_s$ on $\mathcal{X}$: for all $x, x' \in [0,1]^d$, $x \preceq_s x' \Leftrightarrow s(x) \leq s(x')$. From a quantitative perspective, the accuracy of any scoring rule, can be evaluated through the ROC curve criterion, namely the PP-plot $t \in \mathbb{R} \mapsto (1 - H_s(t),\ 1 - G_s(t))$, where $H_s(t) = \mathbb{P}\{s(x) \leq t \mid Y = 0\}$ 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 $\alpha \in (0,1)$. 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 $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $x \in [0,1]^d$. We have $\mathcal{S}^* = \{s \in \mathcal{S} :\ \ \forall x,\ x' \in [0,1]^d,\ \ \eta(x) < \eta(x') \Rightarrow s^*(x) < s^*(x')\}$ 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 $d_\infty(s, s^*) := \sup_{\alpha \in (0,1)}\{\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(s, \alpha)\}$. An alternative convention to represent the ROC of a scoring function s, which we will use for the remainder of this paper, is to consider the broken line $\widetilde{\mathrm{ROC}}(s, .)$, which arises from connecting the PP-plot by line segments at each possible jump of the cdf H_s. **From here on out when referring to the ROC of a scoring function s, we refer to the broken line $\widetilde{\mathrm{ROC}}(s, .)$.**

The active learning setting Whereas in the batch mode, the construction of a nearly optimal scoring function (*i.e.* a function $s \in \mathcal{S}$ such that $d_\infty(s, s^*)$ is 'small' with high probability) is based on a collection of independent training examples given in advance, the objective of an *active learner* is to formulate queries in order to recover sequentially the optimal preorder on the feature space $\mathcal{X}$ defined by the supposedly unknown function η. That is, the active learner plays a game with multiple time steps, where, at time each step $t > 0$, they must choose a point $a_t \in \mathcal{X}$ to query, so as to observe the random label $Y_n \sim \mathcal{B}er(\eta(a_t))$ and refine the scoring model incrementally. After a sufficient number of rounds has elapsed, chosen at the learner's discretion, a final scoring function $\hat{\eta}$, is output.

Assumptions on the feature space and posterior We assume $\mathcal{X}$ to be of the form $[a, b]^d$ for $a, b \in \mathbb{R}$ with $a < b$. For clarity we will assume $\mathcal{X} = [0,1]^d$ from here on out. We assume that the regression function η is β-Holder smooth in each dimension, i.e. Assumption 1.

Assumption 1. *For some $\beta > 0$, known to the learner, there exists a constant $C > 0$ such that $\forall, x, y \in [0,1]^d$,*

$$\forall i \leq d, |\eta(x_i) - \eta(y_i)| \leq C|x_i - y_i|^\beta \ .$$

For clarity we assume $C = 1$, otherwise our results and algorithms would change only in the constant terms, which would then depend upon C. We write $\mathcal{B}$ for the set of all problems satisfying the above assumptions, suppressing the dependency upon β and d in the notation.

Policies and fixed confidence regime. We denote the outputted scoring function of the learner $\hat{\eta} \in S$. The way the learner interacts with the environment - i.e. their choice of points to query, how many samples to draw in total and their choice of $\hat{\eta} \in S$, we term the *policy* of the learner. We write $\mathcal{C}$ for the set of all possible policies of the learner. For a policy $\pi \in \mathcal{C}$ and problem $\nu \in \mathcal{B}$ we denote random variable τ^π_ν as the stopping time of policy π. We write $\hat{\eta}^\pi_\nu$ for the scoring function outputted by policy π on problem ν. Where obvious we may drop the dependency on π, ν in the notation, referring to the scoring function outputted by the learner as simply $\hat{\eta}$. We write $\mathbb{P}_{\nu,\pi}$ as the distribution on all samples gathered by a policy π on problem ν. We similarly define $\mathbb{E}_{\nu,\pi}$. For the duration of this paper we will work in the *fixed confidence regime*. For a confidence level $\varepsilon > 0$ and $\delta > 0$, a policy π is said to be $\mathrm{PAC}(\delta, \varepsilon)$ (probably approximately correct), on the class of problems $\mathcal{B}$, if, $\forall \nu \in \mathcal{B}, \mathbb{P}_{\nu,\pi}(d_\infty(\hat{\eta}, \eta) \leq \varepsilon) \geq 1 - \delta$. The goal of the learner is to then obtain a $\mathrm{PAC}(\delta, \varepsilon)$ policy π, such that the expected stopping time in the worst case, $\sup_{\nu \in \mathcal{B}} \mathbb{E}_{\nu,\pi}[\tau^\pi_\nu]$, is minimised.

Problem complexity With the term "problem complexity", we refer to the minimum number of samples the learner can expect to draw, while remaining $\mathrm{PAC}(\varepsilon, \delta)$. It is a quantity that will depend upon the features the problem, i.e. the shape of the posterior η. Our problem complexity will follow from Cheshire et al. (2023) where the authors consider the problem under the assumption η is piece wise constant on a uniform grid of known size K, see assumption 1. In Cheshire et al. (2023), the authors demonstrate that, for a grid point i, a $\mathrm{PAC}(\varepsilon, \delta)$ learner must be able to correctly distinguish $\mu_i > \mu_j$ vs $\mu_j > \mu_i$ for all $j \in [K] : |\mu_i - \mu_k| \leq \Delta_i$, where, $\Delta_i := \max\Big\{x > 0 : \sum_{i \neq j} x\mathbb{I}\big(|\mu_i - \mu_j| \leq x\big) < K\varepsilon p(1 - \mu_i)\Big\}$. Furthermore they show that this gap Δ_i is tight, in the sense that if the learner can correctly distinguish $\mu_i > \mu_j$ vs $\mu_j > \mu_i$ for all pairs $i, j : |\mu_j - \mu_i| \geq \Delta_i$, they are $\mathrm{PAC}(\delta, \varepsilon)$, see Lemmas 2.2 and 2.3 in Cheshire et al. (2023). We adapt the gap Δ_i to our setting as follows, for a point $x \in \mathcal{X}$ define the following gap $\Delta(x)$,

$$\Delta(x) := \min\left\{z > 0 : z\lambda(\{y : \mathbb{I}\big(|\eta(x) - \eta(y)| \leq z\}\big) \geq \varepsilon p(1-\eta(x))\right\} \vee (1-\eta(x)) \ , \tag{2}$$

where $p := \int_{x\in[0,1]^d} \eta(x)\ dx$. The gap $\Delta(x)$ can be seen as the minimum radius of the ball around x, such that in the worst case, miss ranking all points in $\mathcal{B}^d_{\Delta(x)}(x)$, one suffers more than ε regret. We then define the complexity of a point,

$$H(x) := \frac{\Delta(x)^{-d/\beta}}{\mathrm{kl}(\eta(x) - \Delta(x), \eta(x) + \Delta(x))} \ , \tag{3}$$

with the minimum number of samples needed by the learner then of the order, $\int_{x\in[0,1]^d} H(x)\ dx$. Our sample complexity is seen to be well chosen, as it is attained by the upper and lower bounds of Theorems 3.1 and 2 respectively, up to constant and log terms.

2.1 Related literature

Comparison to the discrete setting There is a wide range of literature concerning ranking in the discrete setting - that is, where one has a finite number of actions to rank. To the best of our knowledge the closest work to our own is that of Cheshire et al. (2023), as they consider directly the discrete version of our setting. For completeness we will first mention other related strands of literature. Several works consider the problem of ranking n experts based on there performance across m tasks. Recently, in Pilliat et al. (2024) they consider this problem in the batch setting. Specifically they deal with the task of ranking the rows of a matrix with isotonic columns, from which the learner receives a batch of noisy observations. As pointed out in Cheshire et al. (2023), algorithms designed for the batch setting perform poorly in active bipartite ranking. In Saad et al. (2023) the authors tackle the problem of ranking experts in an active learning setting, wherein the learner makes sequential queries to pairs of actions and experts, receiving a noisy evaluation of an experts performance on said action. They make a monotonictiy assumption on the experts, that is for each pair of experts, one outperforms the other on all tasks. If one restricts to the setting $m = 1$, their setting is very similar to that considered in Cheshire et al. (2023). However, crucially, they only consider the case where the learner is required to return a perfect ranking with high probability. This fundamentally changes the problem when compared to Cheshire et al. (2023) and our own setting, as here the learner is only required to return an essentially "ε good ranking" - that is the the ROC curve of their estimated scoring function must be within epsilon of the optimal ROC curve, in the d_∞ distance.

Another stream of literature concerns pairwise comparisons. In the active setting, see e.g. Jamieson and Nowak (2011), Heckel et al. (2019). In Heckel et al. (2019), each pair of actions $i, j \in [K]$ has a corresponding probability $M_{i,j}$, being the probability item i beats item j. The correct ranking is then given by an ordering according to the Borda scores of the actions. Compared to our setting, differing strategies are required to deal with noisy pairwise comparisons. To estimate the Borda score of a particular action, one must sample many pairs. However, in our setting, one can estimate the score, i.e. the η of an action by sampling it only. Furthermore in Heckel et al. (2019), the goal of the learner is to give a perfect ranking with high probability, which again distances their work from our own. In the Falahatgar et al. (2017) the authors tackle the problem of finding a "ε-good" Borda ranking of the arms, in the sense that the learner cannot incorrectly rank any two arms who's Borda scores differ by more than ε. In this setting, for a certain arm, one needs to ensure that either, it is well ranked against all others with high probability, or that the width of the confidence interval for the Borda score of said arm, is less than ε. In our setting, on the other hand, the required confidence for a single point is more complex. Due to the global nature of the ROC curve, it does not suffice to simply have the level of confidence at ε, across the feature space, but rather the confidence level must depend on local and global properties of the regression function, see Equation (3). For similar reasons, considering a PAC(ε, δ) setup with regards to either the maximum difference between missed ranked arms or the number of incorrectly ranked arms, as in Busa-Fekete et al. (2014), is of a very different nature to our work.

Comparison to $\mathcal{X}$ - armed bandits The $\mathcal{X}$ armed bandit is well studied, particularly for the problems of optimisation and cumulative regret minimisation. As in our case, the majority of the literature makes some form of smoothness assumption on the regression function. For both optimisation and cumulative regret minimisation, it is natural to consider functions that are smooth around one of their global optima. This is in sharp contrast to our setting, as ranking is a global problem, one needs a global smoothness constraint. In Bubeck et al. (2011) the authors propose an algorithm `HOO` which utilises a hierarchical partitioning of the feature space, by iteratively growing a binary tree. When choosing the next node to split, they follow an optimistic policy, choosing the node with the highest upper bound on its expected payoff. Naturally, the expected utility of a node differs fundamentally from the case of optimisation to that of ranking. Further more, such a tree based algorithm would be cumbersome in our setting, as the width of KL divergence based confidence intervals varies across the feature space. The algorithm `HOO` takes the smoothness parameters as known, in Grill et al. (2015), and further developed in Bartlett et al. (2019), the authors consider the problem of optimisation and propose an algorithm `POO`, which takes `HOO` as a subroutine and runs across many smoothness parameters, this cross validation step is an attempt at dealing with unknown smoothness. In Grill et al. (2015) the authors also consider a none topological smoothness assumption, that relates directly to the hierarchical partition of the feature space. Essentially, they assume that in the cell of depth h containing the optima, the regression function is close to the optima on some rate depending on h, this idea is further developed in Bartlett et al. (2019). It would be unsuitable to formulate our problem under such a hierarchical smoothness assumption. As the ranking problem is global, we would need to control the variance of the function in each cell giving something very similar to our Hölder smoothness assumption while being more cumbersome. In Locatelli and Carpentier (2018) it is shown that such an adaptive approach is impossible in the case of cumulative regret. The $\mathcal{X}$ armed bandit has seen some interest beyond cumulative regret and optimisation, in Torossian et al. (2019) they consider quantile and CVaR optimisation. The problem of optimisation for the $\mathcal{X}$ armed bandit has also been considered in the noiseless setting, see Malherbe and Vayatis (2017), Kawaguchi et al. (2016). There are multiple works considering infinite armed bandit problems where there is no topological relation between the arm indices and means, i.e. when the learner draws arms from a reservoir, see Chaudhuri and Kalyanakrishnan (2017),Aziz et al. (2018),(Katz-Samuels and Jamieson, 2020), Heide et al. (2021). The lack of such topological relation gives the above literature a very different flavour to our own.

Novelty of our results in comparison to Cheshire et al. (2023) Let us consider the performance of a naive adaptation of the `active-rank` algorithm of Cheshire et al. (2023) to our setting. The `active-rank` algorithm is designed to operate under the assumption that η is piece wise constant on some uniform grid of known size K. However, under our Hölder smoothness constraint on η, if one runs `active-rank` on a uniform grid of high enough level, we can expect `active-rank` to be PAC(ε, δ) in our setting. For this to be the case, roughly speaking, the discretisation error associated with our proposed adaptation must be at most $\min_{x\in[0,1]} \Delta(x)$. Consequently, for such a naive approach we would need to run the `active-rank` on the grid $\mathcal{E}_d\big(\min_{x\in[0,1]} \Delta(x)^{1/\beta}\big)$, which would then have roughly the following upper bound on expected sampling time, up to log terms,

$$\sum_{i\in\mathcal{E}_d\left(\min\limits_{x\in[0,1]} \Delta(x)^{1/\beta}\right)} \max_{y\in[0,1]} \frac{1}{\mathrm{kl}(\eta(y)-\Delta_i, \eta(y)+\Delta_i)}\ , \tag{4}$$

where we let Δ_i be the maximum of $\Delta(x)$ for x in the ith section of the grid. The above bound can differ considerably from our sample complexity, Equation (3), in the case where the gap $\Delta(x)$ varies across the interval. Such a cases are exactly the ones of interest, where one wishes to exploit the benefits of active learning. Also, to adapt the `active-rank` algorithm in such a way, one would need to have knowledge of $\min_{x\in[0,1]} \Delta(x)$. Our `smooth-rank` is able to vary the level of discretisation across the interval, according to $\Delta(x)$, without requiring any knowledge of $\Delta(x)$ itself. Furthermore, the theoretical bounds in Cheshire et al. (2023) fail to exploit the fact that KL divergence based confidence intervals change in width across the interval $[0, 1]$, being tighter when closer to 0 or 1. This problem is acknowledged and stated as an open question in Cheshire et al. (2023) and our Theorem 3.1 correctly captures this dependency.

3 MAIN THEORETICAL RESULTS

3.1 The `smooth-rank` algorithm

We first establish some notation. We start at time $t = 0$ and each time `smooth-rank` draws a sample we progress to the next time step. At time t, for $i \in [0, 1]^d$ we write $N_t(i)$ for the total number of samples drawn from point i up to time t and $\hat{\mu}_i^t$ for the empirical mean of point i, calculated from all samples drawn from point i up to time t. For exploration parameter $\beta(t, i, \delta) : \mathbb{N} \times [0, 1]^d \times \mathbb{R}_+ \to \mathbb{R}_+$, we then define the LCB and UCB index,

$$\mathrm{LCB}(i, t) := \min\left\{ q \in \big[0, \hat{\mu}_i^t\big] : \mathrm{kl}\big(\hat{\mu}_i^t, q\big) \leq \frac{\beta(t, i, \delta)}{N_t(i)} \right\}, \tag{5}$$

$$\mathrm{UCB}(i, t) := \max\left\{ q \in \big[\hat{\mu}_i^t, 1\big] : \mathrm{kl}\big(\hat{\mu}_i^t, q\big) \leq \frac{\beta(t, i, \delta)}{N_t(i)} \right\}. \tag{6}$$

For the empirical gap at time t at point i, we write, $\hat{\Delta}_{i,t} = \mathrm{UCB}(i, t) - \mathrm{LCB}(i, t)$. For $t > 0$, the algorithm `smooth-rank` maintains an estimate $\widehat{p}_t$ of the proportion p. We write $\hat{\Delta}_{(p),t}$ for the width and upper and lower bounds of the confidence interval centered at $\widehat{p}_t$, calculated as in Equations (5) and (6). The algorithm `smooth-rank` is an elimination algorithm, in that it maintains an active section of the feature space, denoted $S_t \subset [0, 1]^d$ at time t, with $S_0 = [0, 1]^d$ and sequentially eliminates sections of the active set, until time t at which $S_t = \emptyset$. Sections are eliminated as follows, we maintain an active set of points $\mathcal{X}_t \subset [0, 1]^d$, from which `smooth-rank` draws samples. At each time step we sample the point $\arg\max_{i\in\mathcal{X}_t\cap S_t}(\hat{\Delta}_{i,t})$, then, if a point i in $\mathcal{X}_t$ satisfies our elimination rule, see Equation 7, we remove the set $\{x : \arg\min_{j\in\mathcal{X}_t}(\|j - x\|_d) = i\}$ from S_t. At all times we have a natural ordering on $\mathcal{X}_t$ according to the empirical means. Upon termination of the algorithm we will discretise and rank the feature space, according to $\mathcal{X}_t$. Our goal is to only eliminate sections of S_t, once we are confident in their ranking. As we have only a finite number of points in $\mathcal{X}_t$ at any time t, we will suffer a discretisation error. To be as efficient as possible we would like to ensure our discretisation error matches roughly our stochastic error, with this in mind, we ensure that the set of points $\mathcal{X}_t$ provides a covering, up to a certain level $\Delta_{(t)} := \max_{j \in \mathcal{X}_t \cap S_t}(\hat{\Delta}_{j,t})$, of S_t. This is achieved by adding new points to $\mathcal{X}_t$ across the running time, as $\Delta_{(t)}$ decreases with t. In this fashion, `smooth-rank` will vary the final discretisation level across the feature space - the longer a section of the feature space remains in the active set, the greater the ultimate level of disctretisation will be on that section. As the max size of $\mathcal{X}_t$ is not fixed, the exploration parameter will need to grow with the size of $\mathcal{X}_t$ we will set $\beta(t, i, \delta) = c \log(t^2 \hat{\Delta}_{i,t}^{-d/\beta}/\delta)$, with $c > 0$ an absolute constant. This is in contrast to the `active-rank` algorithm of Cheshire et al. (2023), which does not have this problem as they have an upper bound K on the number of points the algorithm will draw samples from.

Elimination rule Roughly speaking, we wish for a point $i \in \mathcal{X}_t$ to satisfy our elimination rule when $\Delta_{(t)} \lesssim \Delta(i)$. To do this we will essentially estimate $\Delta(i)$ and define our elimination rule as follows. At time $t > 0$ for $i \in \mathcal{X}_t$, define, $U_{i,t}(z) := \left\{ j \in \mathcal{X}_t \cap S_t : |\widehat{\mu}_i^t - \widehat{\mu}_j^t| \le z \right\}$ and then define the set of points,

$$\mathcal{Q}_t := \Bigg\{ i \in \mathcal{X}_t \cap S_t : \Delta_{(t)} \le \left(\frac{\varepsilon \hat{p}_t}{\Delta_{(t)}^{d/\beta} |U_{i,t}(6\Delta_{(t)})|} \wedge 1 \right) (1 - \widehat{\mu}_i^t) \Bigg\} . \tag{7}$$

If a point i is in $\mathcal{Q}_t$, the set $\{x : \arg\min_{j \in \mathcal{X}_t}(\|j - x\|_d) = i\}$ is removed from the active set S_t.

Algorithm 1 `smooth-rank`

```
Input: $\varepsilon > 0, \delta > 0, \beta > 0$
Initialise: $S_0 = [0,1]^d$, $\mathcal{X}_0 = \mathcal{E}_d(0.5)$
repeat
  if $\hat{\Delta}_{(p),t} \ge \Delta_{(t)}$ then
    repeat
      Let $p_t$ be a point drawn uniformly from $[0,1]$ and
      update $\hat{p}_t = ((t-1)\hat{p}_{t-1} + p_t)/t$
    until $\hat{\Delta}_{(p),t} \le \max_{i \in \mathcal{X}_t \cap S_t}(\hat{\Delta}_{i,t})$
  else
    Sample point $\arg\max_{i \in \mathcal{X}_t \cap S_t}(\hat{\Delta}_{i,t})$
  end if
  for $i \in S_t \cap \mathcal{X}_t$ do
    if $i \in \mathcal{Q}_t, \max_{i \in \mathcal{X}_t \cap S_t}(\hat{\Delta}_{i,t}) \le \hat{p}_t/4$ then
      $S_t = S_t \setminus \{x : \arg\min_{j \in \mathcal{X}_t}(\|j - x\|_d) = i\}$
    end if
  end for
  Let $\tilde{n} = \min\left(n : 2^{-n} \le \Delta_{(t)}^{1/\beta}\right)$. Add the points
  $\mathcal{E}_d(2^{-\tilde{n}}) \setminus \mathcal{X}_t$ to $\mathcal{X}_t$.
until $S_t = \emptyset$
Let $\hat{\sigma}$ be the permutation sorting $(\widehat{\mu}_i^t)_{i \in \mathcal{X}_t}$ into ascending
order.
Output: $\hat{\eta}(x) = \hat{\sigma}\left(\arg\min_{j \in \mathcal{X}_t}(\|j - x\|_d)\right)$
```

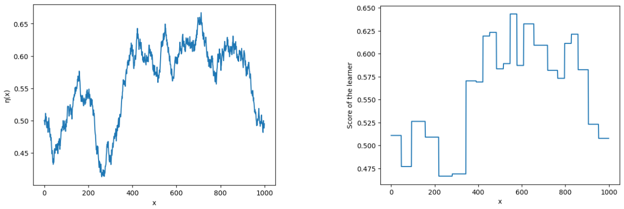

Figure 1: (Left) Regression function, generated by random walk. (Right) Scoring function outputted by `smooth-rank`.

As mentioned, for `smooth-rank`, the final level of discretisation will differ across the feature space. The goal is for the local level of discretisation, to ultimately match the local gap $\Delta(x)$. Otherwise, our discretisation error is either too large, or we sample an unnecessary number of points. As an illustrative example, in Figure 1 we see both the true regression function η and the scoring function outputted by `smooth-rank`. We see the algorithm has a higher level of discretisation where η is more flat, as the gap, $\Delta(x)$ will be smaller on this section. In addition to our varying level of discretisation, our algorithm `smooth-rank` is further distanced from `active-rank` as follows. Instead of pulling all points in the active set at each time step, it samples only the one with maximum gap. This change allow us to capture the true dependence on KL divergence based confidence intervals, an open question stated in Cheshire et al. (2023). In the following Theorem we demonstrate that `smooth-rank` is PAC(ε, δ) and provide an upper bound in its expected sampling time.

Theorem 1. *For $\varepsilon, \delta > 0$, with $\beta(t, i, \delta) = c \log(t^2 \hat{\Delta}_{i,t}^{-d/\beta}/\delta)$ where $c > 0$ is an absolute constant, on all problems $\nu \in \mathcal{B}$, we have that* `smooth-rank` *is PAC(ε, δ), and it's expected sampling time is upper bounded by,*

$$c' \int_{x \in [0,1]^d} H(x) \log(c'' H(x)/\delta) \, dx ,$$

where c', c'' are absolute constants.

Note that the rate exhibited in Theorem depends upon the dimension d, through the sample complexity $H(x)$. The proof of Theorem 3.1 can be split into two parts. The first is to show that the algorithm `smooth-rank` is PAC(ε, δ), see Lemma 4 in the Appendix, the second is to then upper bound the expected stopping time of `smooth-rank`, see Lemma 7 in the Appendix. We will now provide a sketch of proof. All referenced Propositions and Lemmas, can be found in the Appendix, along with their respective proofs.

Proving `smooth-rank` is PAC(ε, δ) To prove `smooth-rank` is PAC(ε, δ), we will consider the favourable event, where at each time step, the true values of η at the sampled points, are within their respective confidence intervals. We define and bound the probability of such an event in the following Lemma.

Lemma 5. *We have that the event,*

$$\mathcal{E} = \bigcap_{t\in\mathbb{N}} \bigcap_{i\in[\mathcal{X}_t]} \left\{\mu_k \in [\mathrm{LCB}(t,i), \mathrm{UCB}(t,i)]\right\} \cap \left\{p \in [\mathrm{LCB}(t,(p)), \mathrm{UCB}(t,(p))]\right\} , \tag{8}$$

occurs with probability greater than $1-\delta$.

The proof of Lemma 5 can be found in Appendix B. While similar in nature to Lemma 3.1 of Cheshire et al. (2023), our Lemma 5 does not follow immediately from a concentration inequality and a union bound, the reason being our exploration parameter grows with the size of the grid. Instead we apply multiple union bounds over many grids, which increase in size geometrically.

To suffer less than ε regret, we know intuitively that, for any given x, the learner must be able to correctly rank the pair x, y for all $y : |\eta(x) - \eta(y)| \le \Delta(x)$. Our careful choice of elimination rule ensures that points are only removed from the active set S_t, when we are confident that this condition is satisfied. The following Lemma demonstrates this property.

Lemma 6. *Upon execution of* `smooth-rank`*, for all* $x \in [0,1]$*, we have that,* $\forall y : |\eta(x) - \eta(y)| \le \Delta(x)$,

$$\mathrm{sign}(\hat{\eta}(x) - \hat{\eta}(y)) = \mathrm{sign}(\eta(x) - \eta(y)) \ .$$

The proof of Lemma 6 follows from Propositions 3 and 4. From this point on the proof follows from analysis of the ROC curve of $\hat{\eta}$ under the condition.

Upper bounding the expected sampling time of `smooth-rank` In the discrete setting, as studied in Cheshire et al. (2023), η is piece wise constant on some uniform grid of known size K and one can then upper bound the total expected number of samples, by upper bounding the expected number of samples on each individual section of the grid. In our continuous setting this strategy is no longer viable, as our level of discretisation will not be fixed across the feature space. However, we can expect certain subsections of the feature space, on which the gap $\Delta(x)$ does not vary too much, to have similar levels of discretisation. To upper bound the expected number of samples drawn by `smooth-rank` on the feature space $[0,1]^d$, we will first upper bound the expected number of total samples on a given subset of the feature space $W \subset [0,1]^d$. We write $N(W)$ for the total number of samples the learner draws on W and $\Delta_W = \min_{x\in W}(\Delta(x))$. In what follows, c, c' are absolute constants which change line by line. First we define the sequence $(t_i)_{i\in\mathbb{N}}$, where

$$t_i := \arg\min \left\{ s : \frac{\log(s^2 \Delta_W^{-d/\beta}/\delta)}{s} \le \max_{x\in W} \mathrm{kl}\big(\eta(x), \eta(x) + 2^{-i}\Delta_W/120\big) \right\} , \tag{9}$$

we suppress the dependence on W in the notation t_i. Roughly speaking once points in W have been sampled t_i times, we can expect them to be removed from the active set with probability converging to 1 as i increases. As the actions of the algorithm are indexed by a global time t, we define T_i as the global time at which a point in W is first sampled t_i times, that is, $T_i := \min(s : \exists j \in \mathcal{X}_s \cap W, N_j(s) \ge t_i)$. We are now ready to define our "good event" $\xi_{W,i}$. Essentially on this event, at time T_i, when a point in W is first sampled t_i times, the empirical means of all active points, $\mathcal{X}_{T_i} \cap S_t$, will be within a distance Δ_W multiplied by a small constant, to their true means, specifically, $\xi_{W,i} := \{\forall j \in \mathcal{X}_{T_i} \cap S_{T_i}, |\widehat{\mu}_j^{T_i} - \eta(j)| \le \Delta_W/120\}$. First we bound the probability of $\xi_{W,i}$,

$$\mathbb{P}(\xi_{W,i}^c) \le c 2^{-i} t_i^{-3} \ , \tag{10}$$

see Proposition 5. We then go on to show that on the event $\xi_{W,i}$, all points in W are eliminated from the active set by time T_i, see Proposition 7. While Proposition 7 uses similar techniques to the proof of Lemma A.6 in Cheshire et al. (2023), these techniques themselves adapted from the fixed confidence best arm identification literature, our event $\xi_{W,i}$ is fundamentally different to that considered in Cheshire et al. (2023). The remaining components of the proof, Propositions 6 through to 15 deal with difficulties specific to the continous setting and have no counterpart in Cheshire et al. (2023).

The total number of samples the learner draws on W does not only depend upon the number of times individual points in W are sampled, but also the number of distinct points sampled in W. Thus, at time T_i we must ensure the discretisation level of the algorithm, $\Delta_{(T_i)}$, is not too small. More precisely, on the event $\xi_{W,i}$ we show that $\Delta_{(T_i)}$ is close to Δ_W, see Proposition 6. We can then go on to show that, on event $\xi_{W,i}$,

$$N(W) \le \lambda(\bar{W}) 2^i \Delta_W^{-d/\beta} t_i \ , \tag{11}$$

where $\bar{W}$, is the completion of W on the grid of level roughly Δ_W, see Proposition 8. We are then finally ready to upper bound $N(W)$,

$$\frac{\mathbb{E}[N(W)]}{\lambda(\bar{W})} \le c \Delta_W^{-d/\beta} \max_{x\in W} H(x) \log(c' H(x)/\delta) \ , \tag{12}$$

see the proof Proposition 10 for the derivation of Equation (12).

From here on we have several technical results, showing that the gap and complexity of points do not differ substantially from W to $\bar{W}$, see Propositions 11 and 12. Finally we consider the sequence of sets, indexed by $n, k \in \mathbb{N}$, as,

$$G_{n,k} = \left\{x : \Delta(x) \in [2^{-n}, 2^{-n-1}], H(x) \in [2^{-k}, 2^{-k-1}]\right\} .$$

We then use (12), to upper bound $\sum_{n,k=1}^{\infty} \mathbb{E}[N(G_{n,k})]$ by

$$c \sum_{n,k \in \lambda(G_{n,k})\mathbb{N}} \max_{x \in G_{n,k}} \Delta(x)^{-\beta} H(x) \log(c' H(x)/\delta) \ ,$$

see Proposition 14. It then remains to lower bound the integral,

$$\int_{y \in [0,1]^d} \frac{\Delta(y)^{-\beta} \log(H(y))}{\mathrm{kl}(\eta(y) - \Delta(y), \eta(y) + \Delta(y))} \, dy \ ,$$

by the above summation.

3.2 Lower bound

In the following theorem we demonstrate a lower bound on the expected sampling time of any PAC(ε, δ) algorithm, for the 1 dimensional case, the proof can be found in section D.

Theorem 2. *Let $\varepsilon \in [0, 1/4), 0 < \delta < 1 - \exp(-1/8), d = 1$ and $\nu \in \mathcal{B}$. For any PAC(ε, δ) policy π, there exists a problem $\bar{\nu} \in \mathcal{B}$ such that, for all $x \in [0, 1]$, $\bar{\Delta}(x) \geq \Delta(x)/2$ where $\bar{\Delta}(x)$ is the gap of point x on problem $\bar{\nu}$, where the expected stopping time of policy π on problem $\bar{\nu}$ is bounded as follows,*

$$c' \int \bar{H}(x) \ ,$$

where $c' > 0$ is an absolute constant and $\bar{H}(x)$ is the complexity of point x on problem $\bar{\nu}$.

3.3 Continuous label fixed threshold setting

In the continuous label fixed threshold setting, the setup is as follows. We no longer restrict the labels Y to take a binary value in $\{0, 1\}$, but rather allow for continuous labels, taking a value in $\mathbb{R}$. That is, when sampling a point $x \in [0, 1]^d$, the learner observes the realisation of some random variable, with distribution function F_x. For a fixed threshold $\rho \in [0, 1]$, known to the learner we can then define the regression function $\eta_\rho(x) = 1 - F_x(\rho) = \mathbb{P}(Y \geq \rho | X = x)$. Our results extend to this setting, see section A where we propose the algorithm `continuous-smooth-rank`, which is similar to `smooth-rank` but makes it's decisions on confidence intervals based on the Dvoretzky–Kiefer–Wolfowitz inequality as opposed to the KL divergence.

4 NUMERICAL EXPERIMENTS

To the best of our knowledge, ranking an $\mathcal{X}$-armed bandit has not been considered in the literature. We can instead compare with discrete methods, adapted to the $\mathcal{X}$-armed setting. The most suitable candidate is the `active-rank` algorithm of Cheshire et al. (2023), as they consider directly the discrete version of our setting and the `active-rank` algorithm is shown to perform well against other potential competitors. The algorithm `active-rank` takes as a parameter K, relating to the piece wise constant assumption of Cheshire et al. (2023), see Equation (1). As in our setting we make no such assumption, we instead pass to the algorithm a fixed K. We test the performance of the algorithms on 1-Hölder smooth regression functions η, generated by random walks. We consider two scenarios, in scenario 1 the random walk will remain constant with probability 0.9 whereas in scenario 2 the random walk jumps at every time step. The result is that, for scenario 2 the gap $\Delta(x)$ will remain constant across large sections of the feature space, as opposed to scenario 2 where it will change more frequently. As a result one may expect algorithms designed for the discrete case to perform better on scenario 1 than 2. This appears to be the case, in Figure 2 we see that `smooth-rank` performs favourably against `active-rank`, especially for small sample times, with the difference being more striking in scenario 2.

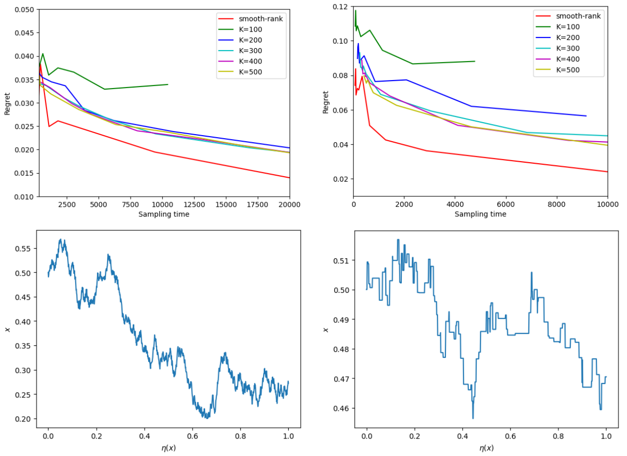

Figure 2: (Top left, top right respectively) Performance of `smooth-rank` compared with `active-rank`, for scenario 1 and scenario 2, for K=(100,200,300,400,500), regret estimated by 50 Monte Carlo realisations of each algorithm. (Bottom left, bottom right respectively) Example simulation of scenario 1 and 2.

4.1 Experiments on simulated credit risk data

Carrying out experiments on real world data, in an active setting has inherent difficulties, i.e. one cannot simply train a classifier on already labeled data. As such, we propose to simulate an active setting, via a real world data set. Credit risk evaluation has been a key motivating application for bipartite ranking in the batch setting. We will use the Home Credit Default Risk Extensive EDA data set, which shows the credit default for over 30000 users, with variety of features. We focus on the features credit and annuity. The regression function being approximated via kernel regression, with the optimal bandwidth chosen via cross validation.

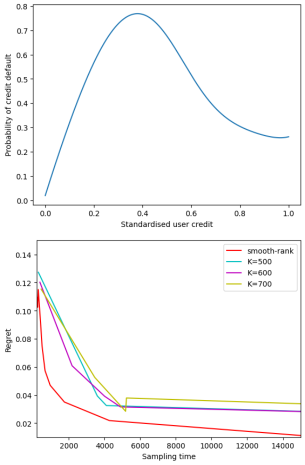

Figure 3: (Top right, top left respectively) Result of KDE for modeling credit default risk given user credit and annuity on EDA data. (Bottom right, bottom left respectively) Performance of `smooth-rank` compared with `active-rank`, for credit default given user credit and annuity, for K=(500,600,700), regret estimated by 50 Monte Carlo realisations of each algorithm.

The performance of our smooth-rank and the active-rank of Cheshire et al. (2023) is then be evaluated for said η. As with our current experiments, we run active-rank for a variety of K, this essentially gives their active-rank algorithm the "benefit of the doubt" and assumes one is somehow able to access the optimal or near optimal value of K. Essentially, this is comparable to our assumed knowledge of β. In Figure 3 we see that `smooth-rank` performs well against active-rank when simulating credit default given user credit. However in the case of simulating credit default given user annuity, active-rank appears to out perform `smooth-rank`, suggesting that care must be taken to ensure simpler, discrete algorithms do not outperform `smooth-rank` in certain practical settings.

Aside from the fact we "only" simulate a real world setting, there are several drawbacks to the above approach. Firstly the connection between the bandwidth of the kernel regression and the smoothness parameter passed to the algorithm is purely heuristic. It would be of interest for a more experimentally dedicated approach to consider practical settings for which the smoothness parameter is known exactly. Secondly we only consider a small subset of the features, this is to keep running time manageable. Ideally one would do feature selection via say PCA, again, we leave this for a more application oriented work. As no attempt was made to minimise constants, they are likely largely overestimated. The constants used in practice differ from their theoretical counterparts.

5 DISCUSSION AND PERSPECTIVES

Adaptation to unknown smoothness Following the work done in optimisation, e.g. Grill et al. (2015) and Bartlett et al. (2019), where the authors consider optimisation of the $\mathcal{X}$-armed bandit under unknown smoothness, a natural question is what happens when the smoothness parameter β is unknown in our setting. However, in optimisation, adaptation to an unknown smoothness parameter is somewhat easier than in ranking. The reason being, one can take a sub routine, which requires a known smoothness parameter and run many such sub routines in parallel for varying degrees of smoothness. The output of the sub routine with best performance can then be chosen, simply by finding the one with highest output. However, in ranking it is not so simple to judge the best ranking rule from a variety of contenders, or more specifically to our setting, judge when a ranking rule guarantees less than epsilon regret. It's therefore of our opinion, that adaptation to smoothness constraints in active ranking, represents a unique and challenging problem, distinct from that faced in continuous armed bandits.

Multi partite and continuous label ranking Our results extend to the case where one observes a continuous label Y, and ranks the feature space according to the probability Y is above some fixed threshold ρ, known to the learner, see section A of the appendix. The case where the threshold ρ is not fixed - instead, for instance, equipped with some prior, remains an open problem. Also of interest is multipartite ranking, in this case the label Y finite number of values. Such a setting, at least in the discrete setting, would become very related to the ranking experts problem. Multipartite ranking would essentially be the ranking experts problem, see Saad et al. (2023), but with the modification that, for each query the learner observes the performance of an expert on a single problem, chosen at random, as apposed to the experts performance on all problems simultaneously.

Acknowledgements

The work of J. Cheshire is supported by the FMJH, ANR-22-EXES-0013.

References

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

Akhavan, A., Pontil, M., and Tsybakov, A. (2020). Exploiting higher order smoothness in derivative-free optimization and continuous bandits. *Advances in Neural Information Processing Systems*, 33:9017–9027.

Aziz, M., Anderton, J., Kaufmann, E., and Aslam, J. (2018). Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*, pages 3–24.

Bartlett, P. L., Gabillon, V., and Valko, M. (2019). A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. In Garivier, A. and Kale, S., editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 184–206. PMLR.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2011). X-armed bandits. *Journal of Machine Learning Research*, 12(5).

Busa-Fekete, R., Szorenyi, B., and Hüllermeier, E. (2014). Pac rank elicitation through adaptive sampling of noisy preferences. page accepted.

Chaudhuri, A. R. and Kalyanakrishnan, S. (2017). Pac identification of a bandit arm relative to a reward quantile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Cheshire, J., Laurent, V., and Clémençon, S. (2023). Active bipartite ranking. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Clémençon, S. and Achab, M. (2017). Ranking data with continuous labels through oriented recursive partitions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

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

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

Clémençon, S., Robbiano, S., and Vayatis, N. (2013). Ranking data with ordinal labels: optimality and pairwise aggregation. *Machine Learning*, 91(1):67–104.

Clémençon, S. and Thomas, A. (2018). Mass volume curves and anomaly ranking. *Electronic Journal of Statistics*, 12(2):2806 – 2872.

Falahatgar, M., Hao, Y., Orlitsky, A., Pichapati, V., and Ravindrakumar, V. (2017). Maxing and ranking with few assumptions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Grill, J.-B., Valko, M., and Munos, R. (2015). Black-box optimization of noisy functions with unknown smoothness. *Advances in Neural Information Processing Systems*, 28:667–675.

Heckel, R., Shah, N. B., Ramchandran, K., and Wainwright, M. J. (2019). Active ranking from pairwise comparisons and when parametric assumptions do not help.

Heide, R. d., Cheshire, J., Menard, P., and Carpentier, A. (2021). Bandits with many optimal arms. In *Advances in Neural Information Processing Systems*.

Jamieson, K. G. and Nowak, R. (2011). Active ranking using pairwise comparisons. *Advances in neural information processing systems*, 24.

Katz-Samuels, J. and Jamieson, K. (2020). The true sample complexity of identifying good arms. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1781–1791. PMLR.

Kawaguchi, K., Maruyama, Y., and Zheng, X. (2016). Global continuous optimization with error bound and fast convergence. *Journal of Artificial Intelligence Research*, 56:153–195.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

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

Malherbe, C. and Vayatis, N. (2017). Global optimization of Lipschitz functions. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2314–2323. PMLR.

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

Pilliat, E., Carpentier, A., and Verzelen, N. (2024). Optimal rates for ranking a permuted isotonic matrix in polynomial time. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 3236–3273. SIAM.

Saad, E. M., Verzelen, N., and Carpentier, A. (2023). Active ranking of experts based on their performances in many tasks. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29490–29513. PMLR.

Torossian, L., Garivier, A., and Picheny, V. (2019). X-armed bandits: Optimizing quantiles and other risks.

Checklist

1. For all models and algorithms presented, check if you include:
   - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
   - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   - (b) Complete proofs of all theoretical results. [Yes]
   - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [No]
   - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [No]
   - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   - (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   - (b) The license information of the assets, if applicable. [Not Applicable]
   - (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   - (d) Information about consent from data providers/curators. [Not Applicable]
   - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Extension of results to continuous label fixed threshold setting

In the continuous label fixed threshold setting, the setup is as follows. We no longer restrict the labels Y to take a binary value in $\{0, 1\}$, but are rather allow for continuous labels, taking a value in $\mathbb{R}$. That is, when sampling a point $x \in [0, 1]^d$, the learner observes the realisation of some random variable, with distribution function F_x. For a fixed threshold $\rho \in [0, 1]$, known to the learner we can then define the regression function $\eta_\rho(x) = 1 - F_x(\rho) = \mathbb{P}(Y \geq \rho | X = x)$. The setting then follows as in the bipartite case, indeed bipartite ranking can be seen as a restricted version of the continuous label setting. Our approach in the continuous label setting, will mirror that for bipartite ranking, the main difference being that our algorithm will base its decisions on confidence bounds constructed via the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality, as opposed to the KL divergence. The DKW inequality controls the probability that an empirical distribution function differs from its true counterpart. Specifically, for some iid random variables $X_1, ..., X_n$ with distribution function F, we can define the empirical distribution,

$$\hat{F}^n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_1 \geq x) \ ,$$

the DKW inequality then states, for $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} |\hat{F}^n(x) - F(x)| \geq \varepsilon \right) \leq c \exp\bigl(-2n\varepsilon^2\bigr) \ .$$

As the DKW is known to be tight in the worst case, it is reasonable to then update our sample complexity for the continuous label case as follows,

$$H(x) = \Delta(x)^{-d/\beta - 2} \ .$$

As one can see, the DKW inequality is not locally dependent, i.e. for a fixed number of samples the width of the generated confidence bound is not dependent on the value of the respective empirical mean, as opposed to KL divergence based confidence intervals which become tighter as one approaches 0 or 1. As a result, our algorithm for the continuous label case is simpler than `smooth-rank`. We no longer sample the point in the active set with largest confidence interval, but rather progress round by round, where at each round all all points in the active set are sampled. As a result, for a given round t, the widths of the confidence bound, on all points in the active will be constant, denoted $\Delta_{(t)}$, where,

$$\Delta_{(t)} := \arg\min \left\{ \Delta \in [0, 1] : \Delta \geq \sqrt{\frac{\log(t^2 \Delta^{-d/\beta} / \delta)}{t}} \right\} .$$

With our new confidence interval in mind, the elimination rule then follows as for `smooth-rank`,

$$\mathcal{Q}_t := \left\{ i \in \mathcal{X}_t \cap S_t : \Delta_{(t)} \leq \left(\frac{\varepsilon \hat{p}}{\Delta_{(t)}^{d/\beta} |U_{i,t}(6\Delta_{(t)})|} \wedge 1 \right) (1 - \widehat{\mu}_i^t) \right\} .$$

At round t, for a point $i \in \mathcal{X}_t$, let $\hat{F}_i^t$ denote the empirical distribution function of the point i, based on all samples collected from point i thus far. As in `smooth-rank`, we will maintain an estimate of the overall proportion, denoted $F_{(p)}$.

Algorithm 2 `continuous-smooth-rank`

```
Input: $\varepsilon, \delta, \beta$
Initialise: $S_0 = [0,1]^d$, $\mathcal{X}_0 = \mathcal{E}_d(0.5)$, $t = 1$
repeat
  Let $x_t$ be a point drawn uniformly from $[0,1]$ and update $\hat{F}_{(p)} = \Big((t-1)\hat{F}_{(p)} + \mathbb{I}(x_t \geq \rho)\Big)/t$,
  Sample all points in $\mathcal{X}_t$ once and update empirical means.
  for $i \in S_t \cap \mathcal{X}_t$ do
    if $i \in \mathcal{Q}_t, \Delta_{(t)} \leq \hat{F}_{(p)}/4$ then
      $S_t = S_t \setminus \{x : \arg\min_{j \in \mathcal{X}_t}(\|j - x\|_d) = i\}$
    end if
  end for
  Let $\tilde{n} = \min\Big(n : 2^{-n} \leq \Delta_{(t)}^{1/\beta}\Big)$. Add the points $\mathcal{E}_d(2^{-\tilde{n}}) \setminus \mathcal{X}_t$ to $\mathcal{X}_t$.
  $t = t + 1$
until $S_t = \emptyset$
Let $\hat{\sigma}$ be the permutation sorting $(\hat{F}_i^t)_{i \leq |\mathcal{X}_t|}$ into ascending order.
Output: $\hat{\eta}(x) = \hat{\sigma}\big(\arg\min_{j \in \mathcal{X}_t}(||j - x||_d)\big)$
```

The following Lemma mirrors Lemma 5, adapted to the continuous label setting,

Lemma 1. *We have that the event,*

$$\mathcal{E} = \bigcap_{t \in \mathbb{N}} \bigcap_{i \in [\mathcal{X}_t]} \{|\hat{F}_i^t(\rho) - F_i(\rho)| \leq \Delta_{(t)}\} \cup \{|\hat{F}_{(p)} - p| \leq \Delta_{(p),t}\} ,$$

occurs with probability greater than $1 - \delta$*.*

Proof. Let $t > 0$ and $i \in \mathcal{X}_t$, via the DKW inequality we have that,

$$\mathbb{P}\Big(|\hat{F}_i^t(\rho) - F_i(\rho)| \leq \Delta_{(t)}\Big) \leq \frac{\delta \Delta_{(t)}^{d/\beta}}{t^2}$$

The proof of Lemma 1 then follows as in the proof of 5. □

With Lemma 1 in mind, as our elimination rule remains unchanged, the proof that `continuous-smooth-rank` is PAC(ε, δ), demonstrated in the following Lemma, follows as in the proof of 4.

Lemma 2. *For* $\varepsilon, \delta > 0$*, on all problems* $\nu \in \mathcal{B}$*, we have that* `continuous-smooth-rank` *is PAC*(ε, δ)*.*

In the following Lemma we demonstrate an upper bound on the expected sampling time of `continuous-smooth-rank`.

Lemma 3. *For* $\varepsilon, \delta > 0$*, on all problems* $\nu \in \mathcal{B}$*, we have that the expected sampling time of* `continuous-smooth-rank` *is upper bounded by,*

$$c' \int_{x \in [0,1]^d} H(x) \log(c'' H(x)/\delta) \, dx \ ,$$

where c'*,* c'' *are absolute constants.*

Proof. To upper bound the expected sampling time of `continuous-smooth-rank` it is straightforward to adapt the proof of 7, with some slight modifications. First we redefine the t_is as follows,

$$t_i = \arg\min \left\{ t : t \geq \frac{\log\Big(2^i t^2 \Delta_{(t)}^{-1/\beta}/\delta\Big)}{\Delta_W^2} \right\}$$

and then consider the slightly reformulated "good event",

$$\xi_{W,t} := \Big\{ \forall j \in \mathcal{X}_t \cap S_t, |\hat{F}_j^t(\rho) - F_j(\rho)| \leq \Delta_W/120 \Big\} \ .$$

For $t \geq t_i$, via the DKW inequality we have that for $i \in \mathcal{X}_t$,

$$\mathbb{P}\Big(|\hat{F}_i^t(\rho) - F_i(\rho)| \leq \Delta_W/120\Big) \leq \exp\big(-ct\Delta_W^2\big) \leq \delta t^{-2} 2^{-i} \Delta_{(t)}^{-1/\beta} \ .$$

and thus via a union bound, $\mathbb{P}(\xi_{W,t}^c) \leq \delta t^{-2} 2^{-i}$. This matches the result of Proposition 5. What remains is to upper bound t_0 similarly and the proof then follows as in the proof of Lemma 7. □

B Proof that `smooth-rank` is PAC(ε, δ)

In this section we will prove the following Lemma.

Lemma 4. *For $\varepsilon, \delta > 0$, with $\beta(t, i, \delta) = c\log(t^2 \hat{\Delta}_{i,t}^{-d/\beta}/\delta)$ where c is an absolute constant, on all problems $\nu \in \mathcal{B}$, we have that* `smooth-rank` *is PAC(ε, δ).*

The proof of Lemma 4 will follow from several sub lemmas and propositions. Firstly we bound the probability of a good event, where all empirical means generated across the run time of the algorithm, are close to their true means.

Lemma 5. *We have that the event,*

$$\mathcal{E} = \bigcap_{t \in \mathbb{N}} \bigcap_{i \in [\mathcal{X}_t]} \{\mu_k \in [\mathrm{LCB}(t,i), \mathrm{UCB}(t,i)]\} \cap \{p \in [\mathrm{LCB}(t,(p)), \mathrm{UCB}(t,(p))]\ ,$$

occurs with probability greater than $1 - \delta$.

Proof. Let $T_{(i)}$ be the time points for which the set of points $\mathcal{X}_t$ is of size between 2^i and 2^{i+1}, that is,

$$T_{(i)} = \{t : 2^i \leq |\mathcal{X}_t| \leq 2^{i+1}\} \ ,$$

let $\mathcal{X}_{(i)}$ be the set of points, that are sampled on points of time in $T_{(i)}$, that is,

$$\mathcal{X}_{(i)} := \{i : \exists t \in T_{(i)} : a_t = i\} \ .$$

We can rewrite event $\mathcal{E}$ as,

$$\bigcap_{i \in \mathbb{N}} \bigcap_{j \in \mathcal{X}_{(i)}} \bigcap_{t \in \{\mathbb{N} : \exists s \in T_{(i)} : N_j(s) = t\}} \{\mu_k \in [\mathrm{LCB}(t,i), \mathrm{UCB}(t,i)]\} \ .$$

For all $i \in \mathcal{X}_{(i)}, t \in T_{(i)}$, we have that, $\hat{\Delta}_{i,t}^{d/\beta} \leq 2^i$, and thus, via Chernoff and the choice of exploration parameter, for $i \in \mathbb{N}, j \in \mathcal{X}_{(i)}, t \in \{\mathbb{N} : \exists s \in T_{(i)} : N_j(s) = t\}$ we have that,

$$\mathbb{P}(\mu_k \notin [\mathrm{LCB}(t,i), \mathrm{UCB}(t,i)]) \leq 2^{-ci} t^{-2} \delta \ .$$

then, via a union bound

$$\mathbb{P}(\mathcal{E}^c) \leq \sum_{i \in N} \sum_{j \in \mathcal{X}_{(i)}} \sum_{t \in \{\mathbb{N} : \exists s \in T_{(i)} : N_j(s) = t\}} 2^{-ci} t^{-2} \delta \ ,$$

and as $|\mathcal{X}_i| \leq 2^{i+1}$ the result follows via choice of the constant c.

□

Proposition 1. *On event $\mathcal{E}$ we have that, for all times t such that a point is removed from the active set, $2p/3 \leq \hat{p}_t \leq 4p/3$.*

Proof. If a section is removed from the active set at time t, we have that $\Delta_{(t)} \leq \hat{p}/4$. Now, on event $\mathcal{E}$ we have $|p - \hat{p}| \leq \Delta_{(t)}$ which, in combination with the fact $\Delta_{(t)} \leq \hat{p}/4$, implies $\Delta_{(t)} \leq (\widehat{p} + \Delta_{(t)})/3$ and thus $\Delta_{(t)} \leq p/3$.

□

Proposition 2. *On event $\mathcal{E}$, at time t, let $x \in S_t$ be such that $6\Delta_{(t)} \leq \Delta(x)$, we have that, $\forall y \in S_t : |\eta(y) - \eta(x)| \geq \Delta(x)$,*

$$\mathrm{sign}(\hat{\eta}(x) - \hat{\eta}(y)) = \mathrm{sign}(\eta(x) - \eta(y)) \ .$$

Proof. For $y \in S_t$, we have that

$$|\widehat{\mu}^t_{\kappa(y,\tau)} - \eta(y)| \leq 2\Delta_{(t)} \ .$$

The proof then follows from application of the triangle inequality. □

Proposition 3. *On event $\mathcal{E}$, at time t, let $i \in \mathcal{X}_t$ be such that $i \in \mathcal{Q}_t$ and $\{x : |\eta(i) - \eta(x)| \leq \Delta(i)\} \subset S_t$. For all $x : |\eta(x) - \widehat{\mu}^t_i| \leq 4\Delta_{(t)}$, we have that, $\forall y \in S_t : |\eta(y) - \eta(x)| \geq \Delta(x)$,*

$$\mathrm{sign}(\hat{\eta}(x) - \hat{\eta}(y)) = \mathrm{sign}(\eta(x) - \eta(y)) \ .$$

Proof. If $i \in \mathcal{Q}_t$ at time t, we have that,

$$6\Delta_{(t)} \leq \left(\frac{\varepsilon \hat{p}}{\Delta_{(t)}^{d/\beta} |U_{i,t}(6\Delta_{(t)})|} \wedge 1 \right) (1 - \widehat{\mu}^t_i) \ ,$$

thus, as $\hat{\Delta}_{i,t} \leq \Delta_{(t)}$,

$$\begin{aligned} 6\Delta_{(t)} &\leq \frac{\varepsilon \widehat{p}_t (1 - \widehat{\mu}^t_i)}{\Delta_{(t)}^{d/\beta} |U_{i,t}(6\Delta_{(t)})|} \\ &\leq \frac{4\varepsilon p (1 - \widehat{\mu}^t_i)}{3\Delta_{(t)}^{d/\beta} |U_{i,t}(6\Delta_{(t)})|} \\ &\leq \frac{4\varepsilon p (1 - \eta(x)) + 5\Delta_{(t)}}{3\Delta_{(t)}^{d/\beta} |U_{i,t}(6\Delta_{(t)})|} \ . \end{aligned} \tag{13}$$

where the second line comes from Proposition 1. Thus as $U_{x,t}(4\Delta_{(t)})) \subset U_{i,t}(6\Delta_{(t)}))$, we have that, $4\Delta_{(t)} \leq \Delta(x)$ and the proof follows by Proposition 2.

□

Proposition 4. *On event $\mathcal{E}$, at time t, let $x \in S_t$, we have that, $\forall y \in [0, 1] \setminus S_t : |\eta(x) - \eta(y)| \geq \Delta(x)$,*

$$\mathrm{sign}(\hat{\eta}(x) - \hat{\eta}(y)) = \mathrm{sign}(\eta(x) - \eta(y)) \ .$$

Proof. The proof will follow by induction. Assume at time t that $\forall x \in S_{t-1}$ we have that, $\forall y \in [S_t] \setminus S_{t-1} : |\eta(x) - \eta(y)| \geq \Delta(x)$,

$$\mathrm{sign}(\hat{\eta}(x) - \hat{\eta}(y)) = \mathrm{sign}(\eta(x) - \eta(y)) \ .$$

Now let $\tilde{x} \in S_t$, and $y \in [0, 1] \setminus S_t : |\eta(\tilde{x}) - \eta(y)| \geq \Delta(\tilde{x}))$, we must show that, .

$$\mathrm{sign}(\hat{\eta}(\tilde{x}) - \hat{\eta}(y)) = \mathrm{sign}(\eta(\tilde{x}) - \eta(y)) \ .$$

If $y \notin S_{t-1}$ we are done, via the inductive assumption, thus assume $y \in S_{t-1}$. If $\tilde{x} \in U_{j,t-1}(4\Delta_{(t)})$ the proof then follows via Proposition 3, thus assume $\tilde{x} \notin U_{y,t-1}(4\Delta_{(t)})$. We then have that, $|\hat{\eta}(\tilde{x}) - \hat{\eta}(y)| \geq 4\Delta_{(t)}$ and the proof follows. □

The following Lemma now follows from Propositions 3 and 4.

Lemma 6. *Upon execution of* `smooth-rank`*, for all $x \in [0, 1]$, we have that, $\forall y : |\eta(x) - \eta(y)| \leq \Delta(x)$,*

$$\mathrm{sign}(\hat{\eta}(x) - \hat{\eta}(y)) = \mathrm{sign}(\eta(x) - \eta(y)) \ .$$

Proof of Lemma 4. The remainder of the proof now follows as in the proof of Lemma A.3, Cheshire et al. (2023). For completeness we include it here, slightly adapted to or setting. Let $\alpha \in [0, 1]$, define the subset $Z_\alpha \subset [0, 1]^d$ such that, $\mathbb{P}(X \in Z_\alpha | Y = -1) = \alpha$, that is,

$$\frac{\lambda(Z_\alpha)(1 - \kappa(Z_\alpha))}{1 - p} = \alpha \ ,$$

and such that, for some $x_\alpha \in [0,1]$,

$$\forall y : \eta(y) > \eta(x_\alpha), y \in Z_\alpha \;, \qquad \forall j : \eta(y) < \eta(x_\alpha), y \notin Z_\alpha \;. \tag{14}$$

We then have $\mathrm{ROC}^*(\alpha) = \mathbb{P}(X \in Z_\alpha | Y = +1) = \frac{\lambda(Z_\alpha)(\kappa(Z_\alpha))}{p}$. The choice of Z_α is not necessarily unique, and as η me be constant across sections of the feature space, likewise x_α is also not necessarily unique, in this case we take arbitrary Z_α, x_α. Now define the subset $\hat{Z}_\alpha \in [0,1]$ such that,

$$\frac{\lambda(\hat{Z}_\alpha)(1-\kappa(\hat{Z}_\alpha))}{1-p} = \alpha \;,$$

and,

$$\forall x \in \hat{Z}_\alpha, y \notin \hat{Z}_\alpha, s_{\hat{\mathcal{P}}}(x) \geq s_{\hat{\mathcal{P}}}(y) \;,$$

so $\mathrm{ROC}(s,\alpha) = \frac{\lambda(\hat{Z}_\alpha)(\kappa(\hat{Z}_\alpha))}{p}$. Again $\hat{Z}_\alpha$ is not necessarily unique, in which case we choose arbitrarily. Via Lemma 6, we have that, $\forall y \in [0,1] : |\eta(y) - \eta(x_\alpha)| \leq \Delta(x_\alpha)$,

$$\mathrm{sign}(\hat{\eta}(y) - \hat{\eta}(x_\alpha)) = \mathrm{sign}(\eta(y) - \eta(x_\alpha)) \;. \tag{15}$$

Let

$$Z'_\alpha = \{y \in Z_\alpha : |\eta(y) - \eta(x_\alpha)| \leq \Delta(x)\} \;, \qquad \hat{Z}'_\alpha = \{y \in \hat{Z}_\alpha : |\eta(y) - \eta(x_\alpha)| \leq \Delta(x)\} \;.$$

Via Equation 15, we have that,

$$\mathrm{ROC}(\alpha, \eta) - \mathrm{ROC}(\alpha, s_{\hat{\mathcal{P}}}) = \frac{\lambda(Z'_\alpha)\kappa(Z'_\alpha)}{p} - \frac{\lambda(\hat{Z}'_\alpha)\kappa(\hat{Z}'_\alpha)}{p} \;.$$

Before finalising the proof we must lower bound $\lambda(\hat{Z}'_\alpha)$ and $\kappa(\hat{Z}'_\alpha)$. We first lower bound $\kappa(\hat{Z}'_\alpha)$.

$$\kappa(\hat{Z}'_\alpha) \geq \eta(x_\alpha) - \Delta_{i_\alpha} \;, \qquad \kappa(Z'_\alpha), \leq \eta(x_\alpha) + \Delta_{i_\alpha} \;. \tag{16}$$

We will now lower bound $\lambda(\hat{Z}'_\alpha)$

$$\frac{\lambda(\hat{Z}'_\alpha)}{\lambda(Z'_\alpha)} = \frac{1-\kappa(Z'_\alpha)}{1-\kappa(\hat{Z}'_\alpha)} \leq \frac{1-\eta(x_\alpha) + \Delta_{i_\alpha}}{1-\eta(x_\alpha) - \Delta_{i_\alpha}} \;. \tag{17}$$

Via combinations of Equations (16) and (17), we have,

$$\begin{aligned}
\frac{\lambda(Z'_\alpha)\kappa(Z'_\alpha)}{p} - \frac{\lambda(\hat{Z}'_\alpha)\kappa(\hat{Z}'_\alpha)}{p} &\leq \frac{1}{p}\Big(\lambda(Z'_\alpha)(\eta(x_\alpha) + \Delta(x_\alpha)) - \lambda(\hat{Z}'_\alpha)(\eta(x_\alpha) - \Delta(x_\alpha))\Big) && (18)\\
&\leq \frac{1}{p}\left(\lambda(Z'_\alpha)(\eta(x_\alpha) + \Delta(x_\alpha)) - \lambda(Z'_\alpha)\frac{(\eta(x_\alpha) - \Delta(x_\alpha))(1-\eta(x_\alpha)+\Delta(x_\alpha)))}{1-\eta(x_\alpha) - \Delta(x_\alpha)}\right) && (19)\\
&\leq \frac{2\lambda(Z'_\alpha)\Delta(x_\alpha)}{p(1-\eta(x_\alpha) - \Delta(x_\alpha))} \leq \frac{2\lambda(Z'_\alpha)\Delta(x_\alpha)}{p(1-\eta(x_\alpha))} && (20)
\end{aligned}$$

It remains to remark that, $\Delta(x_\alpha) \leq \frac{\varepsilon p}{\lambda(Z'_\alpha)}$, by definition, and thus,

$$\mathrm{ROC}(\alpha, \eta) - \mathrm{ROC}\big(\alpha, s_{\hat{\mathcal{P}}}\big) \leq \varepsilon \;.$$

As we chose α w.l.o.g the proof then follows.

□

C Proof of expected sampling time for `smooth-rank`

In this section we will upper bound the expected sampling time of `smooth-rank`, as stated in the following Lemma.

Lemma 7. *For $\varepsilon, \delta > 0$, with $\beta(t, i, \delta) = c \log(t^2 \hat{\Delta}_{i,t}^{-d/\beta}/\delta)$ where c is an absolute constant, on all problems $\nu \in \mathcal{B}$, it's expected sampling time is upper bounded by,*

$$c' \int_{x \in [0,1]^d} H(x) \log(c'' H(x)/\delta) \, dx \ ,$$

where c', c'' are absolute constants.

The proof of the above Lemma 7 will follow from several sub lemmas and propositions. For the entirety of this section, c, c', c'' are absolute constants which vary line by line. Also for clarity, we restrict to the 1 dimensional case, however, *all results generalise immediately to higher dimensions*. To upper bound the expected number of samples the learner draws on the entire feature space, we will first demonstrate an upper bound on a subset of the feature space. Consider a subset $W \subset [0, 1]$. Let $\Delta_W := \min_W(\Delta(x))/4$ be the minimum gap across the subset W, multiplied by some small constant. For our given subset W, we will now define a sequence of integers $t_0, t_1, ...$, here for clarity, we suppress dependency on W in the notation. For $i \in \mathbb{N}$, define,

$$t_i := \arg\min \left\{ s : \frac{\log(s^2 \Delta_W^{-1/\beta}/\delta)}{s} \leq \max_{x \in W} \mathrm{kl}\big(\eta(x), \eta(x) + 2^{-i}\Delta_W/120\big) \right\} .$$

We will show that, once points in W have been sampled t_0 times, they will be eliminated from the active set with constant probability, this probability will converge to 1, as points are sampled $t_1, t_2 ...$ etc. As our notations and the actions of the algorithm are indexed by a global time t, we define T_i as the global time at which a point in W is first sampled t_i times, that is,

$$T_i := \min(s : \exists j \in \mathcal{X}_s, N_j(s) \geq t_i) .$$

We are now ready to define our "good event" $\xi_{W,i}$. Essentially on this event, at time T_i, when a point in W is first sampled t_i times, the empirical means of all active points, $\mathcal{X}_{T_i} \cap S_t$, will be within a distance Δ_W multiplied by a small constant, to their true means, specifically,

$$\xi_{W,i} := \{\forall j \in \mathcal{X}_{T_i} \cap S_{T_i}, |\widehat{\mu}_j^{T_i} - \eta(j)| \leq \Delta_W/120\} .$$

We will go one to show that on the good event $\xi_{W,i}$, all points within W will be eliminated from the active set, see Proposition 7, however, first we upper bound the probability of $\xi_{W,i}^c$ in the following proposition.

Proposition 5. *For $W \subset [0, 1]$, $i \in \mathbb{N}$, we have that,*

$$\mathbb{P}(\xi_{W,i}^c) \leq c 2^{-i} t_i^{-3} .$$

Proof. Set $j = a_{T_i}$, that is the point sampled by the algorithm at time T_i, we have that $N_j(T_i) = t_i$. Now,

$$\mathbb{P}(\mathrm{UCB}(t_i, j) \geq \eta(j) + \Delta_W/120) \leq \mathbb{P}\Big(\mathrm{kl}(\widehat{\mu}_j^{T_i}, \eta(j) + \Delta_W/120) \leq \mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)\Big) . \tag{21}$$

Let,

$$r(\gamma) = \{z \in (\eta(j), \eta(j) + \Delta_W/120) : \mathrm{kl}(z, \eta(j) + \Delta_W/120) = \mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)/\gamma\} .$$

Consider the function $\phi(z) = \mathrm{kl}(\eta(j) + z, \eta(j) + \Delta_W/120)$, on the interval $[0, \Delta(x)/120]$. We have that ϕ is convex and $\phi(\Delta(x)/120) = 0$. Thus for all $z \in \big[0, \frac{\Delta_W}{120}\big]$,

$$\phi(z) \leq (1 - z) \frac{120 \, \mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)}{\Delta_W} . \tag{22}$$

Now consider Equation (22) evaluated at $z = r(\gamma) - \eta(j)$, giving,

$$\frac{\mathrm{kl}(r(\gamma), \eta(j) + \Delta_W/120)}{\mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)} \leq \frac{120}{\Delta_W}(1 + \eta(j) - r(\gamma)) . \tag{23}$$

By definition of $r(\gamma)$, we have,

$$\mathrm{kl}(r(\gamma), \eta(j) + \Delta_W/120) = \mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)/\gamma \tag{24}$$

A combination of Equations (23) and (24) then leads to,

$$r(\gamma) \geq \eta(j) + \Delta_W\left(\frac{1}{120} - \frac{2}{\gamma}\right) \geq \eta(j) + \frac{\Delta_W}{240}\ ,$$

for $\gamma > 480$. We now have,

$$\begin{aligned}
\mathbb{P}\Big(\ \mathrm{kl}(\widehat{\mu}_j^{T_i}, \eta(j) + \Delta_W/120) \leq \frac{\mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)}{\gamma}\Big)& \\
&= \mathbb{P}\Big(\mathrm{kl}(\widehat{\mu}_j^{T_i}, \eta(j) + \Delta_W/120) \leq \mathrm{kl}(r(\gamma), \eta(j) + \Delta_W/120)\Big) \\
&= \mathbb{P}(\widehat{\mu}_j^{T_i} \geq r(\gamma)) \\
&\leq c\exp(-t_i\,\mathrm{kl}(\eta(j), r(\gamma))) \\
&\leq c\exp\left(-\frac{\gamma\,\mathrm{kl}(\eta(j), r(\gamma))}{\mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)}\log(t_i^2\Delta_W^{-1}/\delta)\right) \\
&\leq c\exp\left(-\frac{\gamma\,\mathrm{kl}\big(\eta(j), \eta(j) + \frac{\Delta_W}{240}\big)}{\mathrm{kl}(\eta(j), \eta(j) + 2^{-i}\Delta_W/120)}\log(t_i^2\Delta_W^{-1}/\delta)\right) \\
&\leq c\exp\Big(-4^i\log(t_i^2\Delta_W^{-1/\beta}/\delta)c_\gamma\Big) \\
&\leq c4^{-i}t_i^{-3}\Delta_W^{-d/\beta}
\end{aligned} \tag{25}$$

where c_γ is a constant depending only on γ. Thus, via combination of Equations (C) and (25),

$$\mathbb{P}(\mathrm{UCB}(t_i, j) \geq \eta(j) + \Delta_W/120) \leq c4^{-i}t_i^{-3}\Delta_W^{-1/\beta}\ ,$$

via similar reasoning we have also that,

$$\mathbb{P}(\mathrm{LCB}(t_{j,i}, j) \leq \mu_j - \Delta_W/120) \leq c4^{-i}t_i^{-3}\Delta_W^{-1/\beta}\ .$$

Now take $k \in \mathcal{X}_{T_i} \cap S_{T_i}$, as point j is sampled at time T_i, we have that, $\hat{\Delta}_{j,T_i} = \max_{i \in \mathcal{X}_{T_i} \cap S_{T_i}}\left(\hat{\Delta}_{i,T_i}\right)$ and thus $\hat{\Delta}_{k,T_i} \geq \hat{\Delta}_{j,T_i}$, leading to,

$$\mathbb{P}(\mathrm{LCB}(T_i, k) \leq \eta(k) - \Delta_W/120) \leq \mathbb{P}(\mathrm{LCB}(T_i, j) \leq \eta(j) - \Delta_W/120)$$

and

$$\mathbb{P}(\mathrm{UCB}(T_i, k) \geq \eta(k) + \Delta_W/120) \leq \mathbb{P}(\mathrm{UCB}(T_i, j) \geq \eta(j) + \Delta_W/120)\ .$$

And thus,

$$\mathbb{P}\Big(|\widehat{\mu}_k^{T_i} - \eta(k)| \leq \Delta_W/120\Big) \leq c4^{-i}t_i^{-3}\Delta_W^{-1/\beta}\ .$$

Before continuing note that, as arm j is pulled at time T_i, we have that $\hat{\Delta}_{j,T_i} = \Delta_{(T_i)}$. We now consider two cases, in the first case assume $\hat{\Delta}_{j,T_i}^{-1/\beta} \geq 2^{-i}\Delta_W^{-1/\beta}$, here via Equation (25) we then have,

$$\mathbb{P}\left(\mathrm{kl}(\widehat{\mu}_j^{T_i}, \eta(j) + \Delta_W/120) \leq \frac{\mathrm{kl}(\eta(j), \eta(j) + \Delta_W/120)}{\gamma}\right) \leq c2^{-i}t_i^{-3}\Delta_{(T_i)}^{-1/\beta}\ . \tag{26}$$

It remains to remark that via the action of the algorithm $|\{\mathcal{X}_{T_i} \cap S_{T_i}\}| \leq \Delta_{(T_i)}^{-d/\beta}$, and thus via a union bound,

$$\mathbb{P}(\xi_{W,i}^c) \leq c2^{-i}t_i^{-3} \tag{27}$$

Now we consider the case where $\hat{\Delta}_{j,T_i} < 2^{-i}\Delta_W^{-d}/120$. Via the choice of the exploration parameter of the algorithm, we have due to Chernoff that,

$$\mathbb{P}(|\widehat{\mu}_j^{T_i} - \eta(j)| \geq \hat{\Delta}_{j,T_i}) \leq t_i^{-3}\hat{\Delta}_{j,T_i}^2 \leq \hat{\Delta}_{j,T_i}2^{-i}$$

In this case via Chernoff we have that,

$$\begin{aligned}\mathbb{P}(|\widehat{\mu}_j^{T_i} - \eta(j)| \geq \Delta_{(W)}/120) &\leq \mathbb{P}\Big(|\widehat{\mu}_j^{T_i} - \eta(j)| \geq \hat{\Delta}_{j,T_i}\Big) \\ &\leq 2^{-i}t_i^{-3}\hat{\Delta}_{j,T_i}\end{aligned}$$

The result then follows as in Equation (26). □

To show that all points in W will be eliminated on our good event, we must demonstrate that on event $\xi_{W,i}$, the discretisation level - that is maximum width of the confidence interval, across all active points, $\Delta_{(T_i)}$, is not *too large*. Specifically we need to ensure that $\Delta_{(T_i)}$ is smaller than Δ_W multiplied by some constant. When we later upper bound the expected number of samples the algorithm draws from W, it will also be necessary to show the level of discretisation is not *too small*. Thus in the following proposition we both upper and lower bound $\Delta_{(T_i)}$ on the event $\xi_{W,i}$.

Proposition 6. *For $i \in \mathbb{N}$, on the event $\xi_{W,i}$ we have that,*

$$c'2^{-i}\Delta_W \leq \Delta_{(T_i)} \leq c2^{-i}\Delta_W$$

Proof. Let $j = a_{T(i)}$. The following inequality holds via basic properties of the KL divergence,

$$\begin{aligned}\mathrm{kl}(\widehat{\mu}_j^{T_i}, \widehat{\mu}_j^{T_i} + c\Delta_W) &\leq \mathrm{kl}(\eta(j) + c\Delta_W, \eta(j) + 2c\Delta_W) \\ &\leq 2\,\mathrm{kl}(\eta(j), \eta(j) + c\Delta_W)\ .\end{aligned}$$

Now we consider the definition of t_i, and apply the above inequality,

$$\begin{aligned}t_i &\geq \log(t_i^2\Delta_W^{-1}/\delta)\max_{j\in W}\mathrm{kl}\big(\eta(j), \eta(j) + 2^{-i}\Delta_W/120\big) \\ &\geq \log(t_i^2\Delta_W^{-1}/\delta)\max_{j\in W}\mathrm{kl}\Big(\widehat{\mu}_j^{T_i} + \Delta_W, \widehat{\mu}_j^{T_i} + \Delta_W + 2^{-i}\Delta_W/120\Big) \\ &\leq c\log(t_i^2\Delta_W^{-1}/\delta)\max_{j\in W}\mathrm{kl}\Big(\widehat{\mu}_j^{T_i}, \widehat{\mu}_j^{T_i} + 2^{-i}\Delta_W/120\Big)\ .\end{aligned} \tag{28}$$

By Definition of $\hat{\Delta}_{j,T_i}$ we also have $\forall z \leq \hat{\Delta}_{j,T_i}$,

$$t_i \leq \log(t_i^2 z^2/\delta)\max_{x\in W}\mathrm{kl}\Big(\widehat{\mu}_j^{T_i}, \widehat{\mu}_j^{T_i} + z/120\Big) \tag{29}$$

Thus via combination of Equations (28) and (29) the proof follows. □

We are now ready to show that on our good event, we will eliminate all points within W. For technical reasons, we are required to condition on the intersection of events $\xi_{W,i-1}$ and $\xi_{W,i}$, to ensure all points are eliminated from W at time T_i.

Proposition 7. *For $t > 0$, $W \subset [0,1]$ such that $\forall x \in W, \Delta(x) \leq 4p$, we have that, on event $\xi_{W,i-1} \cap \xi_{W,i}$,*

$$\forall x \in W \cap S_{T_i}, \kappa(x, T_i) \in \mathcal{Q}_{T_i}\ .$$

Proof. Via Proposition 6 we have that, on event $\xi_{W,i-1}$, for all $x \in W$, $\Delta_{(T_{i-1})} \leq \Delta_W/120 \leq \Delta(x)/120$. Thus via the discretisation action of the algorithm,

$$\max_{x,y\in\mathcal{X}_{T_{i-1}}\cap S_{T_{i-1}}}|x - y| \leq (\Delta(x)/120)^{-1/\beta}\ . \tag{30}$$

Now take $y \in U_{x,T_i}(6\Delta_{(T_i)})$, under event $\xi_{W,i}$, we have via Equation (30),

$$|y - \kappa(y, T_i)| \leq (\Delta(x)/120)^{1/\beta}\ ,$$

and then via Holder smoothness,

$$|\eta(y) - \eta(\kappa(y, T_i))| \leq \Delta(x)/120 \ , \tag{31}$$

via definition of $U_{x,T_i}(6\Delta_{(T_i)})$, we have

$$|\eta(x) - \eta(\kappa(y, T_i))| \leq \Delta(x)/120 \ , \tag{32}$$

and thus via combination of Equations (31) and (32),

$$|\eta(x) - \eta(y)| \leq \Delta(x)/120 + \Delta(x)/120 \leq \Delta(x)/60 \ ,$$

giving,

$$U_{x,T_i}(6\Delta_{(T_i)}) \subset \{y : |\eta(x) - \eta(y)| \leq \Delta(x)\} \ . \tag{33}$$

Now, via definition of $\Delta(x)$,

$$\lambda(\{y : |\eta(x) - \eta(y)| \leq \Delta(x)\}) \leq \frac{p\varepsilon(1 - \eta(x))}{\Delta(x)} \ , \tag{34}$$

and so, via the fact that $\Delta_{(T_i)} \leq \Delta(x)/120$ and combination of Equations (33) and (34),

$$\begin{aligned}
\Delta_{(T_i)} \leq 3\Delta(x)/160 &\leq \frac{3\varepsilon p(1 - \eta(x))}{160\lambda(U_{i,T_i}(6\Delta_{(T_i)}))} \\
&\leq \frac{3\varepsilon p((1 - \widehat{\mu}^t_{\kappa(x,T_i)}) + \Delta_{(T_i)})}{80\lambda(U_{i,T_i}(6\Delta_{(T_i)}))} \\
&\leq \left(\frac{\varepsilon\hat{p}}{\lambda(U_{i,T_i}(6\Delta_{(T_i)}))} \wedge 1\right)(1 - \widehat{\mu}^t_{\kappa(x,T_i)}) \ ,
\end{aligned}$$

where the third inequality comes from the fact $\Delta_{(T_i)} \leq \widehat{p}/4$.

□

For the expected number of samples on a given subset W, we can expect our upper bound to depend upon the size of W, however, this relationship is not completely straightforward. The reason being, that the algorithm maintains a set of active points $\{\mathcal{X}_t \cap S_t\}$, from which it samples. A disjoint set W can be arbitrarily small but contain a large number of points in the active set $\{\mathcal{X}_t \cap S_t\}$. Thus we introduce the following, For $W \in [0, 1]$, let, $\tilde{n} = \arg\min\Big(n : 2^{-n} \leq c\Delta_W^{-1/\beta}\Big)$ and consider the set of cells,

$$\bar{W} := \bigcup_{x \in \mathcal{E}_1(2^{-\tilde{n}}) : \exists z \in W : |x - z| \leq 2^{-\tilde{n}}} \{y : |x - y| \leq 2^{-\tilde{n}}\} \ . \tag{35}$$

The set $\bar{W}$ is essentially the smallest possible extension of W, to a set comprised of the union of cells, on the dyadic grid of level smaller than $\Delta_W^{-1/\beta}$.

Proposition 8. *For $W \in [0, 1]$, such that $\forall x \in W, \Delta(x) \leq 4p$, $i \in \mathbb{N}$, on the event $\xi_{W,i-1} \cap \xi_{W,i}$ we have that*

$$N(W) \leq \lambda(\bar{W})2^i\Delta_W^{-d/\beta}t_i \ ,$$

where $N(W)$ is the total number of samples drawn by the learner on W.

Proof. Via Proposition 7 we have that on event $\xi_{W,i-1} \cap \xi_{W,i}$,

$$\forall j \in \{W \cap \mathcal{X}_{T_i}\}, j \in \mathcal{Q}_{(T_i)} \ ,$$

as a result we have that for all

$$t > T_i, W \cap S_t = \emptyset \ .$$

Thus,

$$N(W) \leq t_i|\{\mathcal{X}_{T_i} \cap W\}| \ .$$

From Proposition 6 we have that on the event $\xi_{W,i-1} \cap \xi_{W,i}$, $\Delta_{(T_i)} \geq c\Delta_W^{-1/\beta} 2^{-i}$, and therefore, each cell of the dyadic grid of level smaller than $c\Delta_W^{-1/\beta} 2^{-i}$ contains at most one point of $\{\mathcal{X}_{T_i} \cap S_t\}$. Thus, the set $\bar{W}$ is a union of cells, each of volume $\Delta_W^{-1/\beta}$ and containing at most one point of the set $\{\mathcal{X}_{T_i} \cap W\}$, thus,

$$|\{\mathcal{X}_{T_i} \cap W\}| \leq \lambda(\bar{W})\Delta_W^{-1/\beta} \ ,$$

and the result follows.

□

To bound in expectation the expected sampling time of the algorithm, on a given subset W, we will make use of the following equality $\mathbb{E}[N(W)] \leq \int_{z=0}^{\infty} \mathbb{P}(N(W) \geq z)\ dz$. A key step of our proof will be to replace the aforementioned integral, with a summation, $c\sum_{i=0}^{\infty} \mathbb{P}(N(W) \geq t_i)$. However, to do this we need to ensure the t_i grow at most geometrically. We show just that in the following Proposition.

Proposition 9. *For all $i \in \mathbb{N}$, we have that $4t_i \geq t_{i+1}$.*

Proof. For all $x \in W$, via definition of $\Delta(x)$, we have that, $\Delta(x) \leq (1-\eta(x))$, thus,

$$\forall x \in W, \eta(x) + 120\Delta_W \leq 1 \ . \tag{36}$$

From Equation (36), we have the following,

$$\max_{x\in W} \mathrm{kl}(\eta(x), \eta(x) + 2^{-i}\Delta_W/120) \leq \mathrm{kl}(\eta(x), \eta(x) + 2^{-i-1}\Delta_W/120)/2 \ , \tag{37}$$

Now assume $t_{i+1} \geq 4t_i$, under this assumption we would have,

$$4t_i \leq \frac{\log(16t_i^2\Delta_W^{-a/\beta}/\delta)}{\mathrm{kl}(\eta(x), \eta(x) + 2^{-i-1}\Delta_W/120)} \leq \frac{\log(t_i^2\Delta_W^{-a/\beta}/\delta)}{4\,\mathrm{kl}(\eta(x), \eta(x) + 2^{-i}\Delta_W/120)}$$

where the second inequality comes from application of Equation (37). Which is a contradiction, according to the definition of t_i. □

We are now ready to upper bound the expected number of samples on a given subset W.

Proposition 10. *For $W \subset [0,1]$, such that $\forall x \in W, \Delta(x) \leq 4p$, let $N(W)$ denote the total number of samples drawn on W. We have the following upper bound on, $N(W)$,*

$$\frac{\mathbb{E}[N(W)]}{\lambda(\bar{W})} \leq c\Delta_W^{-d/\beta} \max_{x\in W} H(x)\log(c'H(x)/\delta) \ .$$

Proof.

$$\begin{aligned}
\frac{\mathbb{E}[N(W)]}{\lambda(\bar{W})} &\leq \int_{z=0}^{\infty} \mathbb{P}\big(N(W)/\lambda(\bar{W}) \geq z\big)\ dz & (38)\\
&\leq t_0\Delta_W^{-d/\beta} + \sum_{i=1}^{\infty} \int_{z=2^i t_i \Delta_W^{-d/\beta}}^{2^{i+1}t_{i+1}\Delta_W^{-d/\beta}} \mathbb{P}\big(N(W)/\lambda(\bar{W}) \geq z\big)\ dz & (39)\\
&\leq t_0\Delta_W^{-d/\beta} + c\sum_{i=1}^{\infty} t_i 2^i \Delta_W^{-d/\beta} \mathbb{P}\Big(N(W)/\lambda(\bar{W}) \geq 2^i t_i \Delta_W^{-d/\beta}\Big) & (40)\\
&\leq t_0\Delta_W^{-d/\beta} + c\Delta_W^{-d/\beta} \sum_{i=1}^{\infty} t_i 2^i \mathbb{P}\big(\xi_{W,i-1}^c \cup \xi_{W,i}^c\big) & (41)\\
&\leq t_0\Delta_W^{-d/\beta} + c\Delta_W^{-d/\beta} \sum_{i=1}^{\infty} t_i 2^i \mathbb{P}\big(\xi_{W,i}^c\big) & (42)
\end{aligned}$$

Where the inequality of Equation (40) follows from Proposition 8 and the inequality of Equation (41) follows from Proposition 9. We now need to upper bound $\Delta_W^{-d/\beta} \sum_{i=1}^{\infty} t_i 2^i \mathbb{P}\big(\xi_{W,t_i}^c\big)$. Via Proposition 5 we have,

$$\Delta_W^{-d/\beta} \sum_{i=1}^{\infty} t_i 2^i \mathbb{P}\big(\xi_{W,i}^c\big) \leq \Delta_W^{-d/\beta} \sum_{t=1}^{\infty} t_i 2^i 2^{-i} t_i^{-3} \leq c \Delta_W^{-d/\beta} \ . \tag{43}$$

Combination of Equations (43) and (42) then leads to,

$$\frac{\mathbb{E}[N(W)]}{\lambda(\bar{W})} \leq c t_0 \Delta_W^{-d/\beta} \ .$$

It now remains to upper bound t_0, set $H_W = \max_{x \in W} \mathrm{kl}(\eta(x), \eta(x) + \Delta_W)$,

$$t_0 \leq H_W \log(c' H_W / \delta \Delta_W) \leq \max_{x \in W} H(x) \log(c' H(x) / \delta \Delta_W)$$

□

Now for a given subset W, we have a upper bound on the expected number of samples, which depends upon the minimum gap and maximum sample complexity across W. What remains is to then divide our feature space $\mathcal{X}$ into areas of similar sample complexity and gap. With this in mind, consider the sequence of sets $G_{1,1}, G_{1,2}, \ldots$ where,

$$G_{n,k} = \big\{x : \Delta(x) \in [2^{-n}, 2^{-n-1}], H(x) \in [2^{-k}, 2^{-k-1}]\big\} \ .$$

The final step of the proof will be to apply Proposition 10, to achieve a tight bound on the expected sampling time on each of the subsets $G_{n,k}$, individually. We can then upper bound the total expected sampling time by a summation across all $n, k \in \mathbb{N}$. However, as the reader will recall from Propisiton 10, for a given subset W our upper bound on the expected sampling time across W, depends upon the expanded subset $\bar{W}$, see equation (35). Therefore, we must ensure that, for a given n, k, the sample complexity and gap across $\bar{G}_{n,k}$ do not differ to greatly from $2^{-n}, 2^{-k}$ respectively. In the following Propositions we show exactly that.

Proposition 11. *For a subset $W \subset [0,1]^d$, for all $x \in \bar{W}$, $\Delta(x) \leq 3\Delta_W$*

Proof. As $x \in \bar{W}$, we have that $\exists z \in W, |x - y| \leq \Delta_W^{-d/\beta}$, and via the property of Holder smoothness, $|\eta(x) - \eta(z)| \leq \Delta_W$. Now,

$$\begin{aligned} \lambda(\{y : |\eta(y) - \eta(z)| \leq \Delta_W\}) &\geq \frac{p\varepsilon(1 - \eta(z))}{\Delta_W} \\ &\geq \frac{p\varepsilon(1 - \eta(x) + \Delta_W)}{2\Delta_W} \\ &\geq \frac{p\varepsilon(1 - \eta(x))}{2\Delta_W} \ . \end{aligned} \tag{44}$$

Where the second inequality comes from the fact that $\Delta_W \leq 1 - \eta(x)$. As,

$$\{y : |\eta(y) - \eta(z)| \leq 2\Delta_W\} \subset \{y : |\eta(y) - \eta(x)| \leq 3\Delta_W\} \ .$$

Now, via definition of $\Delta(x)$, we have that,

$$\forall w \leq \Delta(x), \ \lambda(\{y : |\eta(y) - \eta(x)| \leq w\}) \leq \frac{p\varepsilon(1 - \eta(x))}{w} \ , \tag{45}$$

Thus from Equation (44) we have that,

$$\lambda(\{y : |\eta(y) - \eta(x)| \leq 3\Delta_W\}) \geq \frac{p\varepsilon(1 - \eta(x))}{2\Delta_W} \geq \frac{p\varepsilon(1 - \eta(x))}{3\Delta_W} \ ,$$

and thus $\Delta(x) \leq 3\Delta_W$.

□

Proposition 12. *For a subset $W \subset [0,1]^d$, for all $x \in \bar{W}$, $\Delta(x) \geq \Delta_W$.*

Proof. As $x \in \bar{W}$, we have that $\exists z \in W, |x-y| \leq \Delta_W^{-d/\beta}$, and via the property of Holder smoothness, $|\eta(x)-\eta(z)| \leq \Delta_W$. Assume that $\Delta(x) \leq \Delta_W$, we then have,

$$\begin{aligned}\lambda(\{y : |\eta(y) - \eta(x)| \leq \Delta_W\}) &> \frac{p\varepsilon(1-\eta(x))}{\Delta_W} \\ &> \frac{p\varepsilon(1-\eta(z)+\Delta_W)}{\Delta_W} \\ &> \frac{p\varepsilon(1-\eta(z))}{2\Delta_W} \ . \end{aligned} \tag{46}$$

Where the second inequality comes from the fact that $\Delta_W \leq 1-\eta(z)$. As,

$$\{y : |\eta(y)-\eta(x)| \leq \Delta_W\} \subset \{y : |\eta(y)-\eta(z)| \leq 2\Delta_W\}$$

From Equation (46) we have directly that,

$$\lambda(\{y : |\eta(y)-\eta(z)| \leq 2\Delta_W\}) \geq \frac{p\varepsilon(1-\eta(z))}{2\Delta_W} \ ,$$

which is then a contradiction, via the definition of $\Delta(z)$ - see Equation (45) in the proof of Proposition 11 as $\Delta(z) \geq 2\Delta_W$. □

Proposition 13. *Let $n, k \in \mathbb{N}$, for all $x \in \bar{G}_{n,k}$,*

$$2^{k-4} \leq H(x) \leq 2^{k+4} \ .$$

Proof. Let $W = G_{n,k}$, $x \in \bar{W}$ we have that $\exists z \in W : |x-z| \leq \Delta_W^{-1/\beta}$, and via the property of Holder smoothness, $|\eta(x)-\eta(z)| \leq \Delta_W$. Now,

$$\begin{aligned}\mathrm{kl}(\eta(x), \eta(x)+c\Delta(x)) &\leq \mathrm{kl}(\eta(z)+\Delta_W, \eta(z)+\Delta_W+c\Delta(x)) \\ &\leq 2\,\mathrm{kl}(\eta(z), \eta(z)+c\Delta(x)) & (47)\\ &\leq 2\,\mathrm{kl}(\eta(z), \eta(z)+2c\Delta(z)) & (48)\\ &\leq 4\,\mathrm{kl}(\eta(z), \eta(z)+c\Delta(z)) \ . & (49)\end{aligned}$$

Similarly we have,

$$\mathrm{kl}(\eta(z), \eta(z)+c\Delta(z)) \leq 4\,\mathrm{kl}(\eta(x), \eta(x)+c\Delta(x)) \ .$$

□

Combination of Propositions 13, 12 and 11 leads to the following Proposition.

Proposition 14. *Let $n, k \in \mathbb{N}$, such that $2^{-n} \leq 4p$, for all $x \in \bar{G}_{n,k}$, we have that,*

$$x \in \bigcup_{i=n-1,j=k-1}^{i=n+1,j=k+1} G_{n,k} \ ,$$

and thus,

$$c \sum_{n,k\in\mathbb{N}} \max_{x\in\bar{G}_{n,k}} \Delta(x)^{-\beta} H(x) \leq \sum_{n,k\in\mathbb{N}} \max_{x\in G_{n,k}} \Delta(x)^{-\beta} H(x) \ ,$$

for some absolute constant $c > 0$.

Via combination of Propositions 14 and 10 we have that,

$$\sum_{n,k=1}^{\infty} \mathbb{E}[N(G_{n,k})] \leq c \sum_{n,k\in\lambda(G_{n,k})\mathbb{N}} \max_{x\in G_{n,k}} \Delta(x)^{-\beta} H(x) \log(c' H(x)/\delta)$$

It now remains to upper bound the integral

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

with the following,

$$c\sum_{n\geq\log_2(1/4p),k=1}^{\infty}\lambda(G_{n,k})2^{-n\beta}2^{-k}\log(c'H_W/\delta)\ ,$$

where c, c' are absolute constants. Although we only consider $n \geq \log_2(1/4p)$, our bound follows regardless as a result of the following proposition.

Proposition 15. *We have that,*

$$\lambda(\{x : \Delta(x) \leq 2p\}) \geq 1/2\ .$$

Proof. Let

$$x^* = \max\{x : \Delta(x) \geq \eta(x)\}$$

First assume $\eta(x^*) \geq 2p$. In this case we would have

$$\{x : |\eta(x) - \eta(x^*)| \geq 1/2\},$$

contradicting the definition of $\Delta(x)$, for $\varepsilon < 1/4$. Thus we have that, $\eta(x^*) \leq 2p$. And thus $\lambda\{x : \Delta(x) \leq 2p\} \geq 1/2$. □

D Proof of lower bound

Piece wise β-Holder regression functions The class of problems $\mathcal{B}$ consists of all problems ν such that the regression function η is β-Holder. We define a new class of problems $\tilde{\mathcal{B}}$, for which η is piece wise β-holder continuous, on some known ordered partition of the interval $[0,1]$. Specifically, the class $\tilde{\mathcal{B}}$ consists of the problems, for which there exists a known M sized ordered partition of the interval $\mathcal{P} = \{C_1, ..., C_M\}$, where, for each $m \in [M]$ η is β-Holder on C_m, and furthermore, for all $m, n \in [M] : m > n$, we have that,

$$\forall x \in C_m, y \in C_n, \eta(x) > \eta(y)\ .$$

Again, the ordered partition $\mathcal{P}$ is known by the learner. We will now show that problems of the class $\tilde{\mathcal{B}}$ are strictly easier than that of $\mathcal{B}$. As the class of problems $\mathcal{B}$ allows for any feature space of the form $[a,b]^d$ for $a, b \in \mathbb{R}$ with $a < b$, we immediately have the following result.

Lemma 8. *Let $\delta, \varepsilon > 0$. If there exists a PAC(ε, δ) strategy π such that on all problems $\nu \in \mathcal{B}$ the expected sampling time of strategy π is upper bounded by, $c\int H_\nu(x)\,dx$, for some constant $c > 0$, where $H_\nu(x)$ is the complexity of point x on problem ν, then we have that on all problems $\nu \in \tilde{\mathcal{B}}$, the expected sampling time of π is upper bounded by $c\int H_\nu(x)\,dx$.*

Proof of Theorem 2. The proof will now follow by application of a Fano type inequality on a well chosen set of problems.

Step 1: Constructing our well chosen set of problems As is typical in lower bounds, for a given problem ν, we will wish to find a set of alternate problems, carefully chosen such that the gaps $\Delta(x)$ and therefore complexity, are close on the alternate set to ν. In our setting this is tricky, as the gaps $\Delta(x)$ are dependent on the shape of the regression function η. Modifying η locally can potentially have a global effect on the gaps $\Delta(x)$. To overcome this problem we will consider a well chosen M sized partition the interval $[0,1]$, comprised of the sets $D_0, D_1, ..., D_M$. For each $m \in [M]$ there is a corresponding representative $U_m \in D_m$. The partition is chosen such that on each set D_m the gaps $\Delta(x)$ do not vary too much from $\Delta(U_m)$. Furthermore, the representatives are chosen to be sufficiently far from one another in terms of Δ. In this fashion we can essentially modify η on each set D_m without effecting the value of Δ on the other sets, and as such construct our alternate set of problems.

We define a set of points $U_0, U_1, ...$ recursively as follows, $U_0 = \arg\min_{x\in[0,1]}(\Delta(x))$, for $m \geq 0$ we then define,

$$U_{m+1} = \underset{x\in[0,1]}{\arg\min}\{\Delta(x) : \forall j \leq m, \exists k \in [U_j, x] : |\eta(U_j) - \eta(k)| \geq 3\Delta(k) + 3\Delta(U_j)\}\ ,$$

note that we adopt the convention that, for all $a, b \in [0,1]$, $[a,b] = [b,a]$, thus $[U_j, x]$ is well defined in the case where $x < U_j$. Let M be the largest integer for which U_M exists. Note that the sequence $(\Delta(U_m))_{m>M}$ is monotonically increasing and furthermore, for all $x \in [0,1]$,

$$|\{m \in [M] : \nexists y \in [U_m, x] : |\eta(U_m) - \eta(y)| \leq 3\Delta(U_m) + 3\Delta(y)\}| \leq 2 \; . \tag{50}$$

We then define the corresponding set of groups, $D_0, D_1, ..., D_M$ as follows. For $x \in [0,1]$ let $m, n \in [M]$ be such that, $U_m \leq x \leq U_n$ then set $i^+ = m \vee n$ and $i^- = m \wedge n$. If $\nexists k \in [x, U_{i^+}] : |\eta(U_{i^+}) - \eta(k)| \leq 3\Delta_k + 3\Delta(U_{i^+})$ then $x \in D_{i^+}$, otherwise $i \in D_{i^-}$.

Proposition 16. *For all $m \in [M]$ we have that, $\forall x \in D_m, \Delta(U_m) \leq \Delta(x)$.*

Proof. **Case:1 Assume** $x \geq U_m$ Let n be such that $U_m \leq x \leq U_n$. Via Equation (50), we must have that,

$$\exists k \in [n, x] : |\eta(x) - \eta(k)| \leq 3\Delta(k) + 3\Delta_n \; ,$$

and therefore, if $\Delta(x) < \Delta(U_m)$, then,

$$\arg\min\{\Delta(i) : \forall j \leq m-1, \exists k \in [U_j, i] : |\eta(U_j) - \eta(k)| \geq 3\Delta_k + 3\Delta(U_j)\} \neq U_m$$

which is a contradiction via the definition of U_m.

Case:2 Assume $x \leq U_m$ Let n be such that $U_n \leq x \leq U_m$. Via Equation (50), we must have that,

$$\nexists k \in [x, m] : |\eta(x) - \eta(k)| \leq 3\Delta(k) + 3\Delta(m) \; . \tag{51}$$

Now if we also have,

$$\nexists k \in [n, x] : |\eta(x) - \eta(k)| \leq 3\Delta(k) + 3\Delta(n) \; , \tag{52}$$

then, via combination of Equations (51) and (52),

$$\nexists k \in [n, m] : |\eta(x) - \eta(k)| \leq 3\Delta(k) + 3\Delta(m) \wedge \Delta(n) \; .$$

Without loss of generality assume $\Delta(U_m) \leq \Delta(U_n)$. In this case we have that $m < n$, and,

$$\nexists k \in [U_m, U_n] : |\eta(U_m) - \eta(k)| \geq 3\Delta(k) + 3\Delta(U_m) \; ,$$

a contradiction via the definition of U_n.

□

Proposition 17. *For $m \in [M]$, set $W_m = \{x : |\eta(x) - \eta(U_m)| \leq 3\Delta_{U_m}\}$. We have that,*

$$\lambda(D_m) \leq 21\lambda(W_m) \; .$$

Proof. Firstly, take $x \in D_m$ and consider $J_x := \{y : |\eta(x) - \eta(y)| \leq \Delta(x)\}$.

$$\begin{aligned}
\Delta(x) &\leq \frac{\varepsilon p(1-\eta(x))}{\lambda(J_x)} \\
&\leq \frac{7\varepsilon p(1-\eta(x))}{\lambda(J_x)} - \frac{6\varepsilon p\Delta(x)}{\lambda(J_x)} \\
&\leq \frac{7\varepsilon p((1-\eta(U_m)) + 3\Delta(x) + 3\Delta(U_m))}{\lambda(R_x)} - \frac{6\varepsilon p\Delta(x)}{\lambda(R_x)} \\
&\leq \frac{7\varepsilon p(1-\eta(U_m))}{\lambda(J_x)} + \frac{6\varepsilon p\Delta(x)}{\lambda(J_x)} - \frac{6\varepsilon p\Delta(x)}{\lambda(J_x)}
\end{aligned} \tag{53}$$

where the second inequality comes from the fact that $\Delta(x) \leq 1 - \eta(x)$. From proposition 16, $\Delta(x) \geq \Delta(U_m)$
Now consider the sets, $R_1, R_2, ...$ where for $n \in \mathbb{N}$,

$$R_n := \{x \in D_m : \eta(U_m) + (3 + 3 \cdot 2^n)\Delta(U_m) \leq \eta(x) \leq \eta(U_m) + (3 + 3 \cdot 2^{n+1})\Delta(U_m)\} \; .$$

For $n \in \mathbb{N}$ we will upper bound the size of R_n. Note that $\forall x \in R_n$, $\Delta(x) \geq 2^n\Delta(U_m)$, thus via Equation (53), for all $x \in R_n$, $\lambda(J_x) \leq 7 \cdot 2^{-n}\lambda(W_m)$ and thus $\lambda(R_n) \leq 3 \cdot 7 \cdot 2^{-n}\lambda(W_m)$. The result now follows by summing over all R_n.

□

The proof of the following proposition follows via the same argument as in the proof of Proposition 16.

Proposition 18. *For all $m \in [M]$,*

$$|W_m| \leq c\lambda\{x \in W_m : |\eta(x) - \eta(U_m)| \leq \Delta(U_m)\}$$

We are now ready to construct our set of problems. For $m \in [M]$, define the function,

$$f_m(x) = \begin{cases} \forall x : x \mod 8\Delta(U_m)^\beta \leq 4\Delta(U_m)^\beta, & f(x) = x^{1/\beta}, \\ \forall x : x \mod 8\Delta(U_m)^\beta > 4\Delta(U_m)^\beta, & f(x) = -(x-1)^{1/\beta} + 1 \ . \end{cases}$$

Now let $\mathcal{G}_m = \lfloor \lambda(W_m)\Delta(U_m)^{-\beta} \rfloor$, and define the points,

$$(G_{i,m})_{i \in [\mathcal{G}_m]} = \left(\sum_{n=0}^{m-1} \lambda(W_n) + 8i\Delta(U_m)^\beta \right)_{i \in [\mathcal{G}_m]} .$$

Now consider a family of problems ν^Q indexed by $Q \in \{-1, 1\}^{\sum_{m \in [M]} \mathcal{G}_m}$ and for $m \in [M], i \in [\mathcal{G}_m]$, let $Q_i^m = Q(\sum_{n=1}^{m-1} \mathcal{G}_n + i)$, where the target function η_Q corresponding to ν^Q is defined as follows,

$$\eta_Q(x) = \sum_{m \in M} \sum_{i \in [\mathcal{G}_m]} \mathbb{I}(x \in [G_{i,m}, G_{i+1,m}))\eta(U_m)Q_i^m f_m(x) \ .$$

We see that the peak of each bump Essentially the function η_Q can be split into m segments, each one being a series of "bumps" and "dips", each one beginning and ending at $\eta(U_m)$. The ith feature being a bump if $Q_i = 1$, and a dip if $Q_i = 1$. We see that for the mth segment the max of each bump is $\eta(U_m) + \Delta(U_m)$ and the minimum of each dip is $\eta(U_m) - \Delta(U_m)$. Further more, by our choice of $f_m(x)$, we see η_Q is piecewise Hölder smooth. For $m \in [M]$ we define the set of bumps in the mth segment as

$$W_m^{Q,+} = \bigcup_{i \in [\mathcal{G}_m] : Q_i = 1} [G_{i,m}, G_{i+1,m}) \ ,$$

and similarly the set of dips as,

$$W_m^{Q,-} = \bigcup_{i \in [\mathcal{G}_m] : Q_i = -1} [G_{i,m}, G_{i+1,m}) \ .$$

The following Lemma shows that, for a problem $\nu \in \mathcal{B}$, the gaps and complexity across our family problems, ν^Q indexed by Q, does not differ too much from ν.

Lemma 9. *Given, $\nu \in \mathcal{B}$, write $\Delta_Q(x)$ for the gap of point x on problem ν_Q. We have that for all $x \in [0, 1]$, $\Delta(x) \geq c\Delta_Q(x)$, where $c > 0$ is an absolute constant.*

Proof. If $\nexists m : x \in W_m$, we have that $\eta_Q(x) = 0$ and thus $\Delta_Q(x) = 1$. Thus assume $\exists m : x \in W_m$. As $x \in W_m$ we have $\Delta(x) \geq \Delta(U_m)$. Via proposition 18 and identical reasoning to that in Equation (53) we have that $\Delta_Q(x) \geq c\Delta(U_m)$ and the result follows.

□

Step 2: showing that one suffers ε regret on a well chosen event We remind the reader that we denote the scoring function outputted by the learner as $\hat{s}$. For a set $C \subset [0, 1]$, define,

$$\kappa(C) := \frac{1}{\lambda(C)} \int_C \eta(x) \ , dx \ ,$$

and define

$$z_m := \min \left(z : H_{\hat{s}}(z) \geq \frac{\lambda\Big(\bigcup_{n=1}^{m-1} W_n \cup W_m^{Q,+}\Big)(1 - \kappa\Big(\bigcup_{n=1}^{m-1} W_n \cup W_m^{Q,+}\Big)}{(1-p)} \right) ,$$

Let $\hat{Z}_m \subset [0,1]$ be the largest set such that, $\forall x \in \hat{Z}_m, y \notin \hat{Z}_m, \hat{s}(x) \leq \hat{s}(y)$, and,

$$\lambda(\hat{Z}_m)(1-\kappa(\hat{Z}_m)) \leq \lambda\left(\bigcup_{n=1}^{m-1} W_n \cup W_m^{Q,+}\right)(1-\kappa\left(\bigcup_{n=1}^{m-1} W_n \cup W_m^{Q,+}\right) .$$

that is, we have that $\hat{Z}_m = \{x : \hat{s}(x) \geq z_m\}$. Note that $\hat{Z}_m$ is not necessarily unique, in this case we choose an arbitrary such $\hat{Z}_m$. Furthermore define,

$$\hat{Z}_m^0 = \left\{x \in \hat{Z}_m : x \in \bigcup_{n=1}^{m-1} W_n\right\}, \qquad \hat{Z}_m^1 = \left\{x \in \hat{Z}_m : x \in W_m^{Q,+}\right\} ,$$

and

$$\hat{Z}_m^2 = \left\{x \in \hat{Z}_m : x \in W_m^{Q,-} \cup \bigcup_{n=m+1}^{M} W_n\right\} .$$

Now, define the event,

$$\xi_{i,m} := \left\{\{x \in [G_{i,m}, G_{i+1,m}) : \hat{s}(x) > z_m\} \leq \frac{\Delta(U_m)^\beta}{2}\right\} .$$

and then the events,

$$\mathcal{E}_1^m := \left\{\sum_{i \in \mathcal{G}_m : Q_i = 1} \mathbf{1}(\xi_{i,m}) \leq \frac{|\mathcal{G}_m|}{4}\right\}, \qquad \mathcal{E}_0^m := \left\{\sum_{i \in \mathcal{G}_m : Q_i = -1} \mathbf{1}(\xi_{i,m}) \geq \frac{3|\mathcal{G}_m|}{4}\right\} .$$

And let $\hat{D}_m = \bigcup_{n=1}^{m-1} C_n \setminus \hat{Z}_m^0$. Note that under event $\mathcal{E}_1^m$, we have

$$\lambda\left(\hat{Z}_m^1\right) \leq \frac{3\lambda(W_m^{Q,+})}{4} . \tag{54}$$

Also, via the definition of $\hat{Z}_m$, we have that,

$$\lambda(\hat{Z}_m^1)(1-\kappa(W_m^{Q,+})) + \lambda(\hat{Z}_m^2)(1-\kappa(\hat{Z}_m^2)) = \lambda(W_m^{Q,+})(1-\kappa(W_m^{Q,+})) + \lambda(\hat{D}_m)(1-\kappa(\hat{D}_m)) . \tag{55}$$

which leads to,

$$\lambda(\hat{Z}_m^1)\kappa(\hat{Z}_m^2) + \lambda(\hat{Z}_m^2)\kappa(W_m^{Q,+}) = \lambda(\hat{Z}_m^1) + \lambda(\hat{Z}_m^2) - \lambda(W_m^{Q,+})(1-\kappa(W_m^{Q,+})) + \lambda(\hat{D}_m)(1-\kappa(\hat{D}_m)) . \tag{56}$$

and also in combination with equation 54

$$(\lambda(W_m^{Q,+})/4 + \lambda(\hat{D}_m))(1-\kappa(W_m^{Q,+})) \geq \lambda(\hat{Z}_m^2)\kappa(1-\hat{Z}_m^2) , \tag{57}$$

which also gives $\lambda(\hat{Z}_m^2) \geq \lambda(W_m^{Q,+})/4$.

To complete Step: 2 we now lower bound $d_\infty(\hat{s}, \eta)$ on event $\mathcal{E}_1^m$. Firstly note that,

$$\begin{aligned}\mathrm{ROC}\left(\frac{(1-\hat{Z}_m)\lambda(\hat{Z}_m)}{1-p}, \eta\right) &= \frac{\lambda\left(\bigcup_{n=1}^{m-1} W_n \cup W_m^{Q,+}\right)\kappa\left(\bigcup_{n=1}^{m-1} W_n \cup W_m^{Q,+}\right)}{p} \\ &= \frac{\lambda\left(\bigcup_{n=1}^{m-1} W_n\right)\kappa\left(\bigcup_{n=1}^{m-1} W_n\right)}{p} + \frac{\lambda\left(W_m^{Q,+}\right)\kappa\left(W_m^{Q,+}\right)}{p}\end{aligned}$$

therefore, for a problem $\nu^Q : \sum_{i \in [\mathcal{G}_m]} Q_i^m \geq \mathcal{G}_m \lambda(W_m)/2$, on event $\mathcal{E}_1^m$,

$$
\begin{aligned}
d_\infty(\hat{s},\eta) &= 1/p\left(\lambda(\hat{D}_m)\kappa(\hat{D}_m) + \lambda\big(W_m^{Q,+}\big)\kappa\big(W_m^{Q,+}\big) - \frac{\kappa(W_m^{Q,+})\lambda(\hat{Z}_m^1)}{p} - \kappa(W_m^{Q,-})\lambda(\hat{Z}_m^2)\right)\\
&\geq 1/p\Big(\lambda(\hat{D}_m) + \lambda(W_m^{Q,+}) - \lambda(\hat{Z}_m^1) - \lambda(\hat{Z}_m^2)\Big)\\
&\geq 1/p\Big(\lambda(W_m^{Q,+})/4 + \lambda(\hat{D}_m) - \lambda(\hat{Z}_m^2)\Big)\\
&\geq 1/p\left(\frac{\lambda(\hat{Z}_m^2)(1-\kappa(\hat{Z}_m^2)}{1-\kappa(\hat{D}_m)} - \lambda(\hat{Z}_m^2)\right)\\
&\geq 1/p\left(\frac{\lambda(\hat{Z}_m^2)(\kappa(\hat{D}_m)-\kappa(\hat{Z}_m^2))}{4(1-\kappa(\hat{D}_m)})\right)\\
&\geq 1/p\left(\frac{\lambda(W_m^{Q,+})(\kappa(W_m^{Q,+})-\kappa(W_m^{Q,-}))}{1-\kappa(W_m^{Q,+})}\right) \geq 1/p\left(\frac{\lambda(W_m^{Q,+})\Delta(U_m))}{4(1-\kappa(W_m^{Q,+}))}\right) \geq \varepsilon
\end{aligned}
$$

where the first inequality follows from Equation (56) the second from (54) and the third from (57).

Let $\mathbb{P}_Q$ correspond to the probability under the distribution on all samples collected by strategy π on problem ν^Q. Thus, as we assume policy π is PAC(δ,ε), on all problems ν^Q, we must have that, on all problems $\nu^Q : \sum_{i\in[\mathcal{G}_m]} Q_i^m \geq \mathcal{G}_m\lambda(W_m)/2$,

$$
\mathbb{P}_Q(\mathcal{E}_1^m) \leq \delta\ , \tag{58}
$$

Via similar reasoning we can show that on all problems $\nu^Q : \sum_{i\in[\mathcal{G}_m]} Q_i^m \leq \mathcal{G}_m/2$, we must have that,

$$
\mathbb{P}_Q(\mathcal{E}_0^m) \leq \delta\ . \tag{59}
$$

Step 4: bounding the probability of the sum of $\xi_{i,m}$ Now, for $m\in[M]$, and $\nu^Q : \sum_{i\in[\mathcal{G}_m]} Q_i^m \leq \mathcal{G}_m/2$, via the Azuma hoeffding inequality applied to the martingale,

$$
\sum_{i:[\mathcal{G}_m]\in C_m, Q_i^m=0} [\mathbf{1}(\xi_{i,m}) - \mathbb{P}_Q(\xi_{i,m})]\ ,
$$

we have that,

$$
\mathbb{P}_Q\left(\sum_{i\in[\mathcal{G}_m]:Q_i^m=0} [\mathbf{1}(\xi_{i,m}) - \mathbb{P}_Q(\xi_{i,m})] \geq \mathcal{G}_m \log\left(\frac{1}{1-\delta}\right)\right) \leq 1-\delta\ . \tag{60}
$$

Thus via combination of Equations (59) and (60) we must have that, $\forall Q : \sum_{i\in[\mathcal{G}_m]} Q_i^m \leq \mathcal{G}_m/2$,

$$
\sum_{i\in[\mathcal{G}_m]:Q_i^m=0} \mathbb{P}_Q(\xi_{i,m}) \leq \frac{\mathcal{G}_m}{4} - \mathcal{G}_m \log\left(\frac{1}{1-\delta}\right) \leq \frac{3\mathcal{G}_m}{8}\ , \tag{61}
$$

where the second inequality comes from our assumption $\delta \leq 1-\exp(-1/8)$. Via similar reasoning we also have that, $\forall Q : \sum_{i\in[\mathcal{G}_m]} Q_i^m \geq \mathcal{G}_m\lambda(W_m)/2$,

$$
\sum_{i\in[\mathcal{G}_m]:Q_i^m=0} \mathbb{P}_Q(\xi_{i,m}) \geq \frac{\mathcal{G}_m}{2} - \mathcal{G}_m \log\left(\frac{1}{1-\delta}\right) \geq \frac{5\mathcal{G}_m}{8}\ . \tag{62}
$$

Step 5: applying a Fano type inequality We write $\tau_{m,i}$ for the total number of samples the learner draws on $[G_{i,m}, G_{i+1,m})$ and $\tau_{(m)}$ for the total number of samples the learner draws on the set W_m. Let $Q^{(i)}$ be the transformation of Q that flips the ith coordinate,

$$
Q_a^{(k)} = \begin{cases} Q_a \text{ If } a \neq k, \\ 1-Q_a \text{ If } a = k\,. \end{cases}
$$

Consider the class of problems,

$$\mathfrak{Q}_0 = \left\{ Q : \forall m \in [M],\ \sum_{i:[\mathcal{G}_m]} Q_i^m = \frac{\mathcal{G}_m}{2} - 1 \right\}\ ,$$

and

$$\mathfrak{Q} = \left\{ Q : \forall m \in [M],\ \sum_{i:[\mathcal{G}_m]} Q_i^m = \frac{\mathcal{G}_m}{2} \right\}\ ,$$

and fix $m \in [M]$, we see that, for all $Q \in \mathfrak{Q}_0, i : Q_i = 0$, there exists a unique $\tilde{Q} \in \mathfrak{Q}$ such that $\tilde{Q}^{(i)} = Q$, therefore,

$$\sum_{Q\in\mathfrak{Q}} \sum_{m\in[M]} \sum_{i\in[\mathcal{G}_m]:Q_i=1} \mathbb{P}_{Q^{(i)}}(\xi_{i,m}) = \sum_{Q\in\mathfrak{Q}_0} \sum_{i\in[\mathcal{G}_m]:Q_i=0} \mathbb{P}_Q(\xi_{i,m}) \le \frac{3\mathcal{G}_m}{8}\ , \tag{63}$$

where the final inequality follows from Equation (61). Thus, via combination of Equations (62) and (63), and using the data processing inequality and the convexity of the relative entropy we have,

$$\begin{aligned}
&\mathrm{kl}\left(\underbrace{\frac{1}{|\mathfrak{Q}|} \sum_{Q\in\mathfrak{Q}} \sum_{m\in[M]} \frac{2}{\mathcal{G}_m} \sum_{i\in\mathcal{G}_m:Q_i=1} \mathbb{P}_{Q^{(i)}}(\xi_{i,m})}_{\le 6/8}, \underbrace{\frac{1}{|\mathfrak{Q}|} \sum_{Q\in\mathfrak{Q}} \frac{2}{\mathcal{G}_m} \sum_{i\in\mathcal{G}_m:Q_i=1} \mathbb{P}_Q(\xi_{i,m})}_{\ge 10/8} \right) \\
&\le \frac{1}{|\mathfrak{Q}|} \sum_{Q\in\mathfrak{Q}} \frac{2}{\mathcal{G}_m} \sum_{i\in\mathcal{G}_m:Q_i=1} \mathbb{E}_Q[\tau_{m,i}] \frac{\mathrm{kl}(\eta(U_m) - 4/3\Delta(U_m), \eta(U_m) + 4/3\Delta(U_m))}{2} \\
&\le \frac{1}{|\mathfrak{Q}|} \sum_{Q\in\mathfrak{Q}} \frac{\mathbb{E}_Q[\tau_{(m)}]\,\mathrm{kl}(\eta(U_m) - 4/3\Delta_{U_m}, \eta(U_m) + 4/3\Delta(U_m))}{\mathcal{G}_m}\ .
\end{aligned}$$

Then using the Pinsker inequality $\mathrm{kl}(x, y) \ge 2(x - y)^2$, we obtain

$$\frac{10}{8} \le \frac{3}{8} + \sqrt{\le \frac{1}{|\mathfrak{Q}|} \sum_{Q\in\mathfrak{Q}} \frac{\mathbb{E}_Q[\tau_{(m)}]\,\mathrm{kl}(\eta(U_m) - 4/3\Delta(U_m), \eta(U_m) + 4/3\Delta(U_m))}{\mathcal{G}_m}}\ ,$$

and therefore,

$$\le \frac{1}{|\mathfrak{Q}|} \sum_{Q\in\mathfrak{Q}} \frac{\mathbb{E}_Q[\tau_{(m)}]\,\mathrm{kl}(\eta(U_m) - \Delta(U_m), \eta(U_m) + \Delta(U_m))}{\mathcal{G}_m} \ge \frac{9}{64}$$

thus

$$\begin{aligned}
\le \frac{1}{|\mathfrak{Q}|} \sum_{Q\in\mathfrak{Q}} \mathbb{E}_Q[\tau_{(m)}] \ge\ & c' \frac{\mathcal{G}_m}{\mathrm{kl}(\eta(U_m) - 4/3\Delta(U_m), \eta(U_m) + 4/3\Delta(U_m))} && (64) \\
\ge\ & c' \frac{\lambda(W_m)\Delta(U_m)^\beta}{\mathrm{kl}(\eta(U_m) - 4/3\Delta(U_m), \eta(U_m) + 4/3\Delta(U_m))} && (65)
\end{aligned}$$

and thus,

$$\max_{Q\in\mathfrak{Q}} \sum_{m\in[M]} \mathbb{E}_Q[\tau_{(m)}] \ge\ c' \sum_{m\in[M]} \frac{\lambda(W_m)\Delta(U_m)^\beta}{\mathrm{kl}(\eta(U_m) - 4/3\Delta(U_m), \eta(U_m) + 4/3\Delta(U_m))}$$

where $c' > 0$ is an absolute constant changing line by line. The proof now follows, as $\forall m \in [M], x \in W_m$,

$$H(x) \ge \frac{\Delta(U_m)^\beta}{\mathrm{kl}(\eta(U_m) - 4/3\Delta(U_m), \eta(U_m) + 4/3\Delta(U_m))}\ .$$

□