Online Learning with Abstention

Corinna Cortes [1] **Giulia DeSalvo** [1] **Claudio Gentile** [1 2] **Mehryar Mohri** [3 1] **Scott Yang** [*] [4]

Abstract

We present an extensive study of a key problem in online learning where the learner can opt to abstain from making a prediction, at a certain cost. In the adversarial setting, we show how existing online algorithms and guarantees can be adapted to this problem. In the stochastic setting, we first point out a bias problem that limits the straightforward extension of algorithms such as UCB-N to this context. Next, we give a new algorithm, UCB-GT, that exploits historical data and time-varying feedback graphs. We show that this algorithm benefits from more favorable regret guarantees than a natural extension of UCB-N. We further report the results of a series of experiments demonstrating that UCB-GT largely outperforms that extension of UCB-N, as well as other standard baselines.

1. Introduction

We consider an online learning scenario, prevalent in many applications, where the learner is granted the option of abstaining from making a prediction, at a certain cost. For example, in the classification setting, at each round, the learner can choose to make a prediction and incur a standard zero-one misclassification cost, or elect to abstain, in which case she incurs an abstention cost, typically less than one. Abstention can thus represent an attractive option to avoid a higher cost of misclassification. Note, however, that when the learner abstains, she does not receive the true label (correct class), which results in a loss of information.

This scenario of online learning with abstention is relevant to many real-life problems. As an example, consider the scenario where a doctor can choose to make a diagnosis based on the current information available about a patient, or abstain and request further laboratory tests, which can represent both a time delay and a financial cost. In this case, the abstention cost is usually substantially lower than that of a wrong diagnosis. The online model is appropriate since it captures the gradual experience a doctor gains by testing, examining and following new patients.

Another instance of this problem appears in the design of spoken-dialog applications such as those in modern personal assistants. Each time the user asks a question, the assistant can either offer a direct response to the question, at the risk of providing an inaccurate response, or choose to say "I am sorry, I do not understand?", which results in a longer and thereby more costly dialog requesting the user to reformulate his question. Similar online learning problems arise in the context of self-driving cars where, at each instant, the assistant must determine whether to continue steering the car or return the control to the driver. Online learning with abstention also naturally models many problems arising in electronic commerce platforms such as an Ad Exchange, an online platform set up by a publisher where several advertisers bid in order to compete for an ad slot, the abstention cost being the opportunity loss of not bidding for a specific ad slot.

In the batch setting, the problem of learning with abstention has been studied in a number of publications, starting with (Chow, 1957; 1970). Its theoretical aspects have been analyzed by several authors in the last decade. El-Yaniv & Wiener (2010; 2011) studied the trade-off between the coverage and accuracy of classifiers. Bartlett & Wegkamp (2008) introduced a loss function including explicitly an abstention cost and gave a consistency analysis of a surrogate loss that they used to derive an algorithm. More recently, Cortes et al. (2016a;b) presented a comprehensive study of the problem, including an analysis of the properties of a corresponding abstention (or rejection) loss with a series of theoretical guarantees and algorithmic results both for learning with kernel-based hypotheses and for boosting.

This paper presents an extensive study of the problem of online learning with abstention, in both the adversarial and the stochastic settings. We consider the common scenario of prediction with expert advice (Littlestone & Warmuth, 1994) and adopt the same general abstention loss function as in (Cortes et al., 2016a), with each expert formed by a

[*] Work done at the Courant Institute of Mathematical Sciences. [1] Google Research, New York, NY. [2] INRIA Lille Nord Europe. [3] Courant Institute of Mathematical Sciences, New York, NY. [4] D. E. Shaw & Co., New York, NY. Correspondence to: Giulia DeSalvo <giuliad@google.com>.

Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

pair made of a predictor and an abstention function.

A key aspect of the problem we investigate, which makes it distinct from both batch learning with abstention, where labels are known for all training points, and standard online learning (in the full information setting) is the following: if the algorithm abstains from making a prediction for the input point received at a given round, the true label of that point is not revealed. As a result, the loss of the experts that would have instead made a prediction on that point cannot be determined at that round. Thus, we are dealing with an online learning scenario with partial feedback. If the algorithm chooses to predict, then the true label is revealed and the losses of all experts, including abstaining ones, are known. But, if the algorithm elects to abstain, then only the losses of the abstaining experts are known, all of them being equal to the same abstention cost.

As we shall see, our learning problem can be cast as a specific instance of online learning with a feedback graph, a framework introduced by Mannor & Shamir (2011) and later extensively analyzed by several authors (Caron et al., 2012; Alon et al., 2013; 2014; 2015; Kocák et al., 2014; Neu, 2015; Cohen et al., 2016)). In our context, the feedback graph varies over time, a scenario for which most of the existing algorithms and analyses (specifically, in the stochastic setting) do not readily apply. Our setting is distinct from the KWIK (knows what it knows) framework of Li et al. (2008) and its later extensions, though there are some connections, as discussed in Appendix A.

Our contribution can be summarized as follows. In Section 3, we analyze an adversarial setting both in the case of a finite family of experts and that of an infinite family. We show that the problem of learning with abstention can be cast as that of online learning with a time-varying feedback graph tailored to the problem. In the finite case, we show how ideas from (Alon et al., 2014; 2015) can be extended and combined with this time-varying feedback graph to devise an algorithm, EXP3-ABS, that benefits from favorable guarantees. In turn, EXP3-ABS is used as a subroutine for the infinite case where we show how a surrogate loss function can be carefully designed for the abstention loss, while maintaining the same partial observability. We use the structure of this loss function to extend CONTEXTUALEXP3 (Cesa-Bianchi et al., 2017) to the abstention scenario and prove regret guarantees for its performance.

In Section 4, we shift our attention to the stochastic setting. Stochastic bandits with a fixed feedback graph have been previously studied by Caron et al. (2012) and Cohen et al. (2016). We first show that an immediate extension of these algorithms to the time-varying graphs in the abstention scenario faces a technical *bias problem* in the estimation of the expert losses. Next, we characterize a set of feedback graphs that can circumvent this bias problem in the general setting of online learning with feedback graphs. We further design a new algorithm, UCB-GT, whose feedback graph is *estimated* based on past observations. We prove that the algorithm admits more favorable regret guarantees than the UCB-N algorithm (Caron et al., 2012). Finally, in Section 5 we report the results of several experiments with both artificial and real-world datasets demonstrating that UCB-GT in practice significantly outperforms an unbiased, but limited, extension of UCB-N, as well as a standard bandit baseline, like UCB (Auer et al., 2002a).

2. Learning Problem

Let $\mathcal{X}$ denote the input space (e.g., $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$). We denote by $\mathcal{H}$ a family of predictors $h\colon \mathcal{X} \to \mathbb{R}$, and consider the familiar binary classification problem where the loss $\ell(y, h(x))$ of $h \in \mathcal{H}$ on a labeled pair $(x, y) \in \mathcal{X} \times \{\pm 1\}$ is defined by either the 0/1-loss $1_{yh(x) \leqslant 0}$, or some Lipschitz variant thereof (see Section 3). In all cases, we assume $\ell(\cdot, \cdot) \in [0, 1]$. We also denote by $\mathcal{R}$ a family of abstention functions $r\colon \mathcal{X} \to \mathbb{R}$, with $r(x) \leqslant 0$ indicating an abstention on $x \in \mathcal{X}$ (or that x is *rejected*), and $r(x) > 0$ that x is predicted upon (or that x is *accepted*).

We consider a specific online learning scenario whose regime lies between bandit and full information, sometimes referred to as *bandit with side-information* (e.g., Mannor & Shamir (2011); Caron et al. (2012); Alon et al. (2013; 2014; 2015); Kocák et al. (2014); Neu (2015); Cohen et al. (2016)). In our case, the arms are pairs made of a predictor function h and an abstention function r in a given family $\mathcal{E} \subseteq \mathcal{H} \times \mathcal{R}$. We will denote by $\xi_j = (h_j, r_j)$, $j \in [K]$, the elements of $\mathcal{E}$. In fact, depending on the setting, K may be finite or (uncountably) infinite. Given h_j, one natural choice for the associated abstention function r_j is a confidence-based abstention function of the form $r_j(x) = |h_j(x)| - \theta$, for some threshold $\theta > 0$. Yet, more general pairs (h_j, r_j) can be considered here. This provides an important degree of flexibility in the design of algorithms where abstentions are allowed, as shown in (Cortes et al., 2016a;b). Appendix A presents a concrete example illustrating the benefits of learning with these pair of functions.

The online learning protocol is described as follows. The set $\mathcal{E}$ is known to the learning algorithm beforehand. At each round $t \in [T]$, the online algorithm receives an input $x_t \in \mathcal{X}$ and chooses (possibly at random) an arm (henceforth also called "expert" or "pair") $\xi_{I_t} = (h_{I_t}, r_{I_t}) \in \mathcal{E}$. If the inequality $r_{I_t}(x_t) \leqslant 0$ holds, then the algorithm abstains and incurs as loss an abstention cost $c(x_t) \in [0, 1]$. Otherwise, it predicts based on the sign of $h_{I_t}(x_t)$, receives the true label $y_t \in \{\pm 1\}$, and incurs the loss $\ell(y_t, h_{I_t}(x_t))$. Thus, the overall *abstention loss* L of expert $\xi = (h, r) \in \mathcal{E}$ on the labeled pair $z = (x, y) \in \mathcal{X} \times \{\pm 1\}$ is defined as

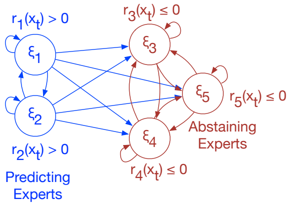

Figure 1: Feedback graph G_t^{ABS} for the scenario of online learning with abstention, with $K = 5$.

follows:

$$L(\xi, z) = \ell(y, h(x))1_{r(x)>0} + c(x)1_{r(x)\leqslant 0}\,. \qquad (1)$$

For simplicity, we will assume throughout that the abstention cost $c(x)$ is a (known) constant $c \in [0, 1]$, independent of x, though all our results can be straightforwardly extended to the case when c is a (Lipschitz) function of x, which is indeed desirable in some applications.

Our problem can be naturally cast as an online learning problem with side information in the form of a feedback graph. Online learning with a feedback graph is a general framework that covers a variety of problems with partial information, including the full information scenario, where the graph is fully connected, and the bandit scenario where all vertices admit only self-loops and are disconnected (Alon et al., 2013; 2014). In our case, we have a directed graph $G_t^{\text{ABS}} = (V, E_t)$ that depends on the instance x_t received by the algorithm at round $t \in [T]$. Here, V denotes the finite set of vertices of this graph, which, in the case of a finite set of arms, coincides with the set of experts $\mathcal{E}$, while E_t denotes the set of directed edges at round t. The directed edge $\xi_i \to \xi_j$ is in E_t if the loss of expert $\xi_j \in V$ is observed when expert ξ_i is selected by the algorithm at round t. In our problem, if the learner chooses to predict at round t (i.e., if $r_{I_t}(x_t) > 0$), then she observes the loss $L(\xi_j, z_t)$ of all experts ξ_j, since the label y_t is revealed to her. If instead she abstains at round t (i.e., if $r_{I_t}(x_t) \leqslant 0$), then she only observes $L(\xi_j, z_t)$ for those experts ξ_j that are abstaining in that round, that is, the set of j such that $r_j(x_t) \leqslant 0$, since for all such ξ_j, we have $L(\xi_j, z_t) = c$. Notice that in both cases the learner can observe the loss of her own action. Thus, the feedback graph we are operating with is a nearly fully connected directed graph with self-loops, except that it admits only one-way edges from predicting to abstaining vertices (see Figure 1 for an example). Observe also that the feedback graph G_t^{ABS} is fully determined by x_t.

We will consider both an adversarial setting (Section 3), where no distributional assumption is made about the sequence $z_t = (x_t, y_t)$, $t \in [T]$, and a stochastic setting (Section 4), where z_t is assumed to be drawn i.i.d. from some unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{\pm 1\}$. For both settings, we measure the performance of an algorithm $\mathcal{A}$ by its *(pseudo-)regret* $R_T(\mathcal{A})$, defined as $R_T(\mathcal{A}) = \sup_{\xi\in\mathcal{E}} \mathbb{E}[\sum_{t=1}^T L(\xi_{I_t}, z_t) - \sum_{t=1}^T L(\xi, z_t)]$, where the expectation is taken both with respect to the algorithm's choice of actions I_ts and, in the stochastic setting, the random draw of the z_ts.

In the stochastic setting, we will be mainly concerned with the case where $\mathcal{E}$ is a finite set of experts $\mathcal{E} = \{\xi_1, \ldots, \xi_K\}$. We then denote by μ_j the expected loss of expert $\xi_j \in \mathcal{E}$, $\mu_j = \mathbb{E}_{z\sim\mathcal{D}}[L(\xi_j, z)]$, by μ^* the expected loss of the best expert, $\mu^* = \min_{j\in[K]} \mu_j$, and by Δ_j the loss gap to the best, $\Delta_j = \mu_j - \mu^*$. In the adversarial setting, we will analyze both the finite and infinite expert scenarios. In the infinite case, since L is non-convex in the relevant parameters (Eq. (1)), further care is needed.

3. Adversarial setting

As a warm-up, we start with the adversarial setting with finitely-many experts. Following ideas from Alon et al. (2014; 2015), we design an online algorithm for the abstention scenario by combining standard finite-arm bandit algorithms, like EXP3 (Auer et al., 2003), with the feedback graph G_t^{ABS} of Section 2. We call the resulting algorithm EXP3-ABS (EXP3 with abstention). The algorithm is a variant of EXP3 where the importance weighting scheme to achieve unbiased loss estimates is based on the probability of the loss of an expert being *observed* as opposed to that of an expert being selected — see Appendix B (Algorithm 3). The following guarantee holds for this algorithm.

Theorem 1 *Let* EXP3-ABS *be run with learning rate η over a set of K experts $\xi_1, \ldots, \xi_K$. Then, the algorithm admits the following regret guarantee after T rounds:*

$$R_T(\text{EXP3-ABS}) \leqslant (\log K)/\eta + \eta T(c^2 + 1)/2.$$

In particular, if EXP3-ABS *is run with $\eta = \sqrt{\frac{2\log K}{(c^2+1)T}}$, then $R_T(\text{EXP3-ABS}) \leqslant \sqrt{2(c^2+1)T\log K}$.*

The proof of this result, as well as all other proofs, is given in the appendix. The dependency of the bound on the number of experts is clearly more favorable than the standard bound for EXP3 ($\sqrt{\log K}$ instead of $\sqrt{K}$). Theorem 1 is in fact reminiscent of what one can achieve using the contextual-bandit algorithm EXP4 (Auer et al., 2002b) run on K experts, each one having two actions.

We now turn our attention to the case of an uncountably infinite $\mathcal{E}$. To model this more general framework, one might be tempted to focus on parametric classes of functions h and r, e.g., the family $\mathcal{E}$ of linear functions

$$\big\{(h, r) : h(x) = w^\top x, r(x) = |w^\top x| - \theta, w \in \mathbb{R}^d, \theta > 0\big\},$$

introduce some convex surrogate of the abstention loss (1), and work in the parametric space of (w, θ) through some

Bandit Convex Optimization technique (e.g., (Hazan, 2016)). Unfortunately, this approach is not easy to put in place, since the surrogate loss not only needs to ensure convexity and some form of calibration, but also the ability for the algorithm to observe the loss of its own action (the self-loops in the graph of Figure 1).

We have been unable to get around this problem by just resorting to convex surrogate losses (and we strongly suspect that it is not possible), and in what follows we instead introduce a surrogate abstention loss which is Lipschitz but not convex. Moreover, we take the more general viewpoint of competing with pairs (h, r) of Lipschitz functions with bounded Lipschitz constant. Let us then consider the version of the abstention loss (1) with $\ell(y, h(x)) = f_\gamma(-yh(x))$, where f_γ is the 0/1-loss with slope $1/(2\gamma)$ at the origin, $f_\gamma(a) = \left(\frac{\gamma+a}{2\gamma}\right) 1_{|a|\leqslant\gamma} + 1_{a\geqslant 0}1_{|a|>\gamma}$ (see Figure 2 (a)), and the class of experts $\mathcal{E} = \left\{\xi = (h, r) \,|\, h, r \colon \mathcal{X} \subseteq \mathbb{R}^d \to [-1, 1]\right\}$. Here, functions h and r in the definition of $\mathcal{E}$ are assumed to be $L_\mathcal{E}$-Lipschitz with respect to an appropriate distance on $\mathbb{R}^d$, for some constant $L_\mathcal{E}$ which determines the size of the family $\mathcal{E}$.

Using ideas from (Cesa-Bianchi et al., 2017), we present an algorithm that approximates the action space by a finite cover while using the structure of the abstention setting. The crux of the problem is to define a Lipschitz function $\widetilde{L}$ that uppers bounds the abstention loss while maintaining the same feedback assumptions, namely the feedback graph given in Figure 1. One Lipschitz function $\widetilde{L}$ that precisely solves this problem is the following:

$$\widetilde{L}(\xi, z) = \begin{cases} c & \text{if } r(x) \leqslant -\gamma \\ 1 + \left(\frac{1-c}{\gamma}\right) r(x) & \text{if } r(x) \in (-\gamma, 0) \\ 1 - \left(\frac{1-f_\gamma(-yh(x))}{\gamma}\right) r(x) & \text{if } r(x) \in [0, \gamma) \\ f_\gamma(-yh(x)) & \text{if } r(x) \geqslant \gamma\ , \end{cases}$$

for $\gamma \in (0, 1)$. $\widetilde{L}(\xi, z)$ is plotted in Figure 2(b). Notice that this function is consistent with the feedback requirements of Section 2: $r_{I_t}(x_t) \leqslant 0$ implies that $\widetilde{L}((h(x_t), r(x_t)), z_t)$ is known to the algorithm (i.e., is independent of y_t) for all $(h, r) \in \mathcal{E}$ such that $r(x_t) \leqslant 0$, while $r_{I_t}(x_t) > 0$ gives complete knowledge of $\widetilde{L}((h(x_t), r(x_t)), z_t)$ for all $(h, r) \in \mathcal{E}$, since y_t is observed.

We can then adapt the machinery from (Cesa-Bianchi et al., 2017) so as to apply a contextual version of EXP3-ABS to the sequence of losses $\widetilde{L}(\xi, z_t), t \in [T]$. The algorithm adaptively covers $\mathcal{X}$ with balls of a fixed radius ε, each ball hosting an instance of EXP3-ABS. We call this algorithm CONTEXP3-ABS – see Appendix B.2 for details.

Theorem 2 *Consider the abstention loss*

$$L(\xi, z) = f_\gamma(-yh(x))1_{r(x)>0} + c1_{r(x)\leqslant 0}\ ,$$

and let $\xi^* = (h^*, r^*) = \operatorname{argmin}_{\xi\in\mathcal{E}} \sum_{t=1}^T L(\xi, z_t)$, with $\mathcal{E} = \{(h, r)\}$ made of pairs of Lipschitz functions as described above. If CONTEXP3-ABS *is run with parameter $\varepsilon \simeq T^{-\frac{1}{2+d}}\gamma^{\frac{2}{2+d}}$ and an appropriate learning rate (see Appendix B), then, it admits the following regret guarantee:*

$$R_T(\text{CONTEXP3-ABS}) \leqslant \widetilde{\mathcal{O}}\left(T^{\frac{d+1}{d+2}}\gamma^{-\frac{d}{d+2}}\right) + M_T^*(\gamma),$$

where $M_T^*(\gamma)$ is the number of x_t such that $|r^*(x_t)| \leqslant \gamma$.

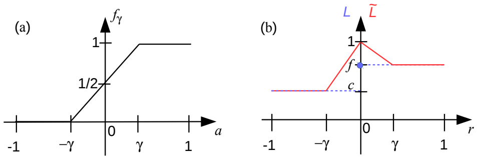

Figure 2: (a) The 0/1-loss function with slope $1/(2\gamma)$ at the origin. (b) For a given value of x and margin $a = -yh(x)$ (which in turn sets the value of $f = f_\gamma(a) \in [0, 1]$), plots of the abstention loss function $L(a, r)$ (dotted blue curve), and the surrogate abstention loss $\widetilde{L}(a, r)$ (red curve), both as a function of $r = r(x) \in [-1, 1]$.

In the above, $\widetilde{\mathcal{O}}$ hides constant and $\ln(T)$ factors, while $\simeq$ disregards constants like $L_\mathcal{E}$, and various log factors. CONTEXP3-ABS is also computationally efficient, thereby providing a compelling solution to the infinite armed case of online learning with abstention.

4. Stochastic setting

We now turn to studying the stochastic setting. As pointed out in Section 2, the problem can be cast as an instance of online learning with time-varying feedback graphs G_t^{ABS}. Thus, a natural method for tackling the problem would be to extend existing algorithms designed for the stochastic setting with feedback graphs to our abstention scenario (Cohen et al., 2016; Caron et al., 2012). We cannot benefit from the algorithm of Cohen et al. (2016) in our scenario. This is because at the heart of its design and theoretical guarantees lies the assumption that the graphs and losses are *independent*. The dependency of the feedback graphs on the observations z_t, which also define the losses, is precisely a property that we wish to exploit in our scenario.

An alternative is to extend the UCB-N algorithm of Caron et al. (2012), for which the authors provide gap-based regret guarantees. This algorithm is defined for a stochastic setting with an undirected feedback graph that is fixed over time. The algorithm can be straightforwardly extended to the case of directed time-varying feedback graphs (see Algorithm 1). We will denote that extension by UCB-NT to explicitly differentiate it from UCB-N. Let $N_t(j)$ denote the set of out-neighbors of vertex ξ_j in the directed graph at time t, i.e., the set of vertices ξ_k destinations of an edge from ξ_j. Then, as with UCB-N, the algorithm updates, at

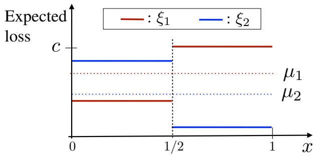

Figure 3: Illustration of the bias problem.

each round t, the upper-confidence bound of every expert for which a feedback is received (those in $N_t(I_t)$), as opposed to updating only the upper-confidence bound of the expert selected, as in the standard UCB of Auer et al. (2002a).

In the context of learning with abstention, the natural feedback graph G_t^{ABS} at time t depends on the observation x_t and varies over time. Can we extend the regret guarantees of Caron et al. (2012) to UCB-NT with such graphs? We will show in Section 4.1 that vanishing regret guarantees do not hold for UCB-NT run with graphs G_t^{ABS}. This is because of a fundamental estimation bias problem that arises when the graph at time t depends on the observation x_t. This issue affects more generally any natural method using the G_t^{ABS} graphs. Nevertheless, we will show in Section 4.2 that UCB-NT does benefit from favorable guarantees, provided the feedback graph G_t^{ABS} it uses at round t is replaced by one that only depends on events up to time $t-1$.

4.1. Bias problem

Assume there are two experts: ξ_1 (red) and ξ_2 (blue) with $\mu_2 < \mu_1$ and $\mathcal{X} = [0, 1]$ (see Figure 3). For $x > \frac{1}{2}$, the red expert ξ_1 is abstaining and incurring a loss c, whereas the blue expert is never abstaining. Assume that the probability mass is quasi-uniform over the interval $[0, 1]$ but with slightly more mass over the region $x < \frac{1}{2}$. The algorithm may then start out by observing points in this region. Here, both experts accept and the algorithm obtains error estimates corresponding to the solid red and blue lines for $x < \frac{1}{2}$. When the algorithm observes a point $x > \frac{1}{2}$, it naturally selects the red abstaining expert since it admits a better current estimated loss. However, for $x > \frac{1}{2}$, the red expert is worse than the blue expert ξ_2. Furthermore, it is abstaining and thus providing no updates for expert ξ_2 (which is instead predicting). Hence, the algorithm continues to maintain an estimate of ξ_2's loss at the level of the blue solid line indicated for $x < \frac{1}{2}$; it then continues to select the red expert for all xs and incurs a high regret.[1]

This simple example shows that, unlike the adversarial scenario (Section 3), G_t^{ABS}, here, cannot depend on the input x_t, and that, in general, the indiscriminate use of feedback graphs may result in biased loss observations. On the other hand, we know that if we were to avoid using feedback graphs at all (which is always possible using UCB), we would always be able to define unbiased loss estimates. A natural question is then: can we construct time-varying feedback graphs that lead to unbiased loss observations? In the next section, we show how to design such a sequence of auxiliary feedback graphs, which in turn allows us to then extend UCB-NT to the setting of time-varying feedback graphs for general loss functions. Under this assumption, we can achieve unbiased empirical estimates of the average losses μ_j of the experts, which will allow us to apply standard concentration bounds in the proof of this algorithm.

ALGORITHM 1: UCB-NT

```
for $t \geqslant 1$ do
    RECEIVE($x_t$);
    $\xi_{I_t} \leftarrow \text{argmin}_{\xi_j \in \mathcal{E}} \left\{\widehat{\mu}_{j,t-1} - S_{j,t-1}\right\}$;
    for $\xi_j \in \mathcal{E}$ do
        $Q_{j,t} \leftarrow \sum_{s=1}^{t} 1_{j \in N_s(I_s)}$ ;
        $S_{j,t} \leftarrow \sqrt{\frac{5 \log t}{Q_{j,t}}}$;
        $\widehat{\mu}_{j,t} \leftarrow \frac{1}{Q_{j,t}} \sum_{s=1}^{t} L(\xi_j, z_s) 1_{j \in N_s(I_s)}$.
    end for
end for
```

4.2. Time-varying graphs for UCB-NT

We now show that UCB-NT benefits from favorable guarantees, so long as the feedback graph G_t^{ABS} it uses at time t depends only on events up to time $t-1$. This extension works for general bounded losses and does not only apply to our specific abstention loss L.

So, let us assume that the feedback graph in round t (and the associated out-neighborhoods $N_t(\cdot)$) in Algorithm 1 only depends on the observed losses $L(\xi_i, z_s)$ and inputs x_s, for $s = 1, \ldots, t-1$, and $i \in [K]$, and let us denote this feedback graph by G_t, so as not to get confused with G_t^{ABS}. Under this assumption, we can derive strong regret guarantees for UCB-NT with time-varying graphs using a newly introduced notion of admissible coverings. For the feedback graph at time G_t, let $\mathcal{C}_t$ be a collection of subsets of V covering G_t, such that $\forall C_t \in \mathcal{C}_t$, $i, j \in C_t$ means that $i \in N_t(j)$ and $j \in N_t(i)$. We denote such a collection an *admissible covering* of G_t. Let $\mathcal{F}_t$ denote the set of all admissible coverings of G_t, and let $\mathcal{F} = \cap_{t=1}^{T} \mathcal{F}_t$, i.e. the collection of *shared admissible coverings* that apply across all time steps. Then by construction, for any $\mathcal{C} \in \mathcal{F}$ and $C \in \mathcal{C}$, $i, j \in C$ means that $i \in N_t(j)$ and $j \in N_t(i)$ for every $t \in [T]$. Note that the definition of $\mathcal{F}$ is equivalent to considering the set of edges that are shared across all G_t, and then considering admissible coverings over the graph induced by these shared edges. Moreover, since $\cup_{i=1}^{K}\{i\} \in \mathcal{F}_t$ for every $t \in [T]$, $\mathcal{F}$ is always non-empty.

[1] For the sake of clarity, we did not introduce specific real values for the expected loss of each expert on each of the half intervals, but that can be done straightforwardly. We have also verified experimentally with such values that the bias problem just pointed out indeed leads to poor regret for UCB-NT.

Theorem 3 *Assume that, for all $t \in [T]$, the feedback graph G_t depends only on information up to time $t-1$. Then, the regret of* UCB-NT *is bounded as follows:*

$$\mathcal{O}\left(\mathbb{E}\left[\min_{\mathcal{C}\in\mathcal{F}} \sum_{C\in\mathcal{C}} \frac{\max_{j\in C}\Delta_j}{\min_{j\in C}\Delta_j^2}\log(T)+K\right]\right).$$

The theorem gives a bound on the regret based on any admissible covering that applies to every feedback graph seen during learning, and the minimum chooses the admissible covering with the smallest regret.

Theorem 3 can be interpreted as an extension of Theorem 2 in Caron et al. (2012) to time-varying feedback graphs. Its proof involves showing that the use of feedback graphs G_t that depend only on information up to $t-1$ can result in unbiased loss estimates, and it considers shared admissible coverings that apply across the sequence of feedback graphs to derive a time-varying bound that leverages the shared updates from the graph.

Moreover, the bound illustrates that if the feedback graphs in a problem admit a shared admissible covering with a small number of elements $|\mathcal{C}| \ll K$ (e.g. if the feedback graphs can be decomposed into a small number of components that are fixed across time) for which $\max_{j\in C}\Delta_j \approx \min_{j\in C}\Delta_j$, then this bound can be up to a factor $\frac{|\mathcal{C}|}{K}$ tighter than the bound guaranteed by the standard UCB algorithm. Moreover, this regret guarantee is always more favorable than that of the standard UCB since the (trivial) admissible covering that splits V into K singletons for all t is always an admissible covering of every G_t. Furthermore, note that if the feedback graph is fixed throughout all rounds and we interpret the doubly-directed edges as edges of an undirected graph G_U, it follows that $\mathcal{F} = \mathcal{F}_t$ for every $t \in [T]$. Thus, we straightforwardly obtain the following result, which is comparable to Theorem 2 in (Caron et al., 2012).

Corollary 1 *If the feedback graph $G_t = G$ is fixed over time, then the guarantee of Theorem 3 is upper-bounded by:*

$$\mathcal{O}\left(\min_{\mathcal{C}} \sum_{C\in\mathcal{C}} \frac{\max_{i\in C}\Delta_i}{\min_{i\in C}\Delta_i^2}\log(T)+K\right),$$

the outer minimum being over all admissible coverings $\mathcal{C}$ of G_U.

Caron et al. (2012) present matching lower bounds for the case of stochastic bandits with a fixed feedback graph. Since we can again design abstention scenarios with fixed feedback graphs, these bounds carry over to our setting.

Now, how can we use the results of this section to design an algorithm for the abstention scenario? The natural feedback graphs we discussed in Section 3 are no longer applicable since G_t^{ABS} depends on x_t. Nevertheless, we will present two solutions to this problem. In Section 4.3, we present a solution with a fixed graph G that closely captures the problem of learning with abstention. Next, in Section 4.4, we will show how to define and leverage a time-varying graph G_t that is estimated based on past observations.

4.3. UCB-N with the subset feedback graph

In this section, we define a *subset feedback graph*, G^{SUB}, that captures the most informative feedback in the problem of learning with abstention and yet is safe in the sense that it does not depend on x_t. The definition of the graph is based on the following simple observation: if the abstention region associated with ξ_i is a subset of that of ξ_j, then, if ξ_i is selected at some round t and is abstaining, so is ξ_j. For an example, see ξ_i and ξ_j in Figure 4 (top). Crucially, this implication holds regardless of the particular input point x_t received in the region of abstention of ξ_i. Thus, the set of vertices of G^{SUB} is $\mathcal{E}$, and G^{SUB} admits an edge from ξ_i to ξ_j, iff $\{x \in \mathcal{X} \colon r_i(x) \leqslant 0\} \subseteq \{x \in \mathcal{X} \colon r_j(x) \leqslant 0\}$. Since G^{SUB} does not vary with time, it trivially verifies the condition of the previous section. Thus, UCB-NT run with G^{SUB} admits the regret guarantees of Theorem 3, where we only need to consider the set of admissible coverings of the fixed graph G^{SUB}.

The example of Section 4.1 illustrated a bias problem in a special case where the feedback graphs G_t were not subgraphs of G^{SUB}. The following result shows more generally that feedback graphs not included in G^{SUB} may result in catastrophic regret behavior.

Proposition 1 *Assume that* UCB-NT *is run with feedback graphs G_t that are not subsets of G^{SUB}. Then, there exists a family of predictors $\mathcal{H}$, a Lipschitz loss function ℓ in (1), and a distribution $\mathcal{D}$ over z_ts for which* UCB-NT *incurs linear regret with arbitrarily high probability.*

The proof of the proposition is given in Appendix C.3. In view of this result, no fixed feedback graph for UCB-NT can be more informative than G^{SUB}. But how can we leverage past observations (up to time $t-1$) to derive a feedback graph that would be more informative than the simple subset graph G^{SUB}? The next section provides a solution based on feedback graphs estimated based on past observations and a new algorithm.

4.4. UCB-GT algorithm

We seek graphs G_t that admit G^{SUB} as a subgraph. We will show how certain types of edges can be safely added to G^{SUB} based on past observations. This leads to a new algorithm, UCB-GT (UCB with estimated time-varying graph), whose pseudocode is given in Algorithm 2. As illustrated by Figure 4, the key idea of UCB-GT is to augment G^{SUB} with edges from ξ_j to ξ_i where the subset property

ALGORITHM 2: UCB-GT

```
for $t \geqslant 1$ do
    RECEIVE($x_t$);
    $\xi_{I_t} \leftarrow \mathrm{argmin}_{\xi_i \in \mathcal{E}} \{\widehat{\mu}_{i,t-1} - S_{i,t-1}\}$,
    where $S_{i,t-1}$ is as in Algorithm 1;
    for $\xi_i \in \mathcal{E}$ do
        if $\widehat{p}^{t-1}_{I_t,i} \leqslant \gamma_{i,t-1}$ then $Q_{i,t} \leftarrow Q_{i,t-1} + 1$;
            if $r_{I_t}(x_t) \leqslant 0 \wedge r_i(x_t) > 0$ then
                $\widehat{\mu}_{i,t} \leftarrow \left(1 - \frac{1}{Q_{i,t}}\right) \widehat{\mu}_{i,t-1}$;        (*)
            else $\widehat{\mu}_{i,t} \leftarrow \frac{L(\xi_i, z_t)}{Q_{i,t}} + \left(1 - \frac{1}{Q_{i,t}}\right) \widehat{\mu}_{i,t-1}$;
        else $Q_{i,t} \leftarrow Q_{i,t-1}$, $\widehat{\mu}_{i,t} \leftarrow \widehat{\mu}_{i,t-1}$ .
    end for
end for
```

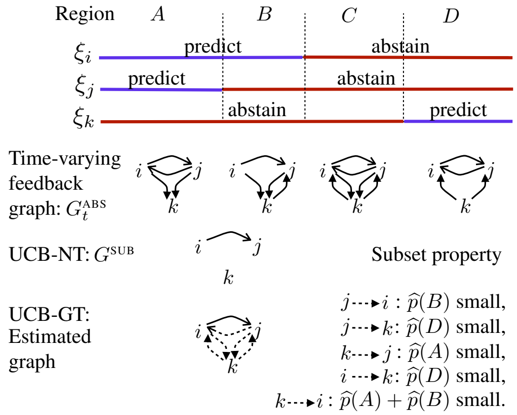
Region
A
B
C
D
ξ_i
predict
abstain
ξ_j
predict
abstain
ξ_k
abstain
predict
Time-varying feedback graph: G_t^{ABS}
i
j
k
UCB-NT: G^{SUB}
Subset property
UCB-GT: Estimated graph
$j \dashrightarrow i$: $\widehat{p}(B)$ small,
$j \dashrightarrow k$: $\widehat{p}(D)$ small,
$k \dashrightarrow j$: $\widehat{p}(A)$ small,
$i \dashrightarrow k$: $\widehat{p}(D)$ small,
$k \dashrightarrow i$: $\widehat{p}(A) + \widehat{p}(B)$ small.

Figure 4: The top row shows three experts ξ_i, ξ_j, and ξ_k on a one-dimensional input space marked by their prediction and abstention regions. Below each region, the time-varying graph G_t^{ABS} is shown. To avoid the bias problem affecting the graphs G_t^{ABS}, one option is to use G^{SUB}. Yet, as illustrated, G^{SUB} is minimal and in this example admits only one edge (excluding self-loops). Thus, a better option is to use the time-varying graphs of UCB-GT since they are richer and more informative. For these graphs, an edge is added from ξ_j to ξ_i if the probability of the region where ξ_j is abstaining but ξ_i is predicting is (estimated to be) small.

$\{x\colon r_j(x) \leqslant 0\} \subseteq \{x\colon r_i(x) \leqslant 0\}$ may not hold, but where the implication $(r_j(x) \leqslant 0 \Rightarrow r_i(x) \leqslant 0)$ holds with high probability over the choice of $x \in \mathcal{X}$, that is, the region $\{x\colon r_j(x) \leqslant 0 \wedge r_i(x) > 0\}$ admits low probability. Of course, adding such an edge $\xi_j \to \xi_i$ can cause the estimation bias of Section 4.1. But, if we restrict ourselves to cases where $p_{j,i} = \mathbb{P}[r_j(x) \leqslant 0 \wedge r_i(x) > 0]$ is upper bounded by some carefully chosen quantity that changes over rounds, the effect of this bias will be limited. In reverse, as illustrated in Figure 4, the resulting feedback graph can be substantially more beneficial since it may have many more edges than G^{SUB}, hence leading to more frequent updates of the experts' losses and more favorable regret guarantees. This benefit is further corroborated by our experimental results (Section 5).

Since we do not have access to $p_{j,i}$, we use instead empirical estimates $\widehat{p}^{t-1}_{j,i} := \frac{1}{t-1} \sum_{s=1}^{t-1} 1_{r_j(x_s) \leqslant 0, r_i(x_s) > 0}$. At time t, if expert ξ_j is selected, we update expert ξ_i if the condition $\widehat{p}^{t-1}_{j,i} \leqslant \gamma_{i,t-1}$ holds with $\gamma_{i,t-1} = \sqrt{5Q_i(t-1)\log(t)/((K-1)(t-1))}$. If the expert ξ_{I_t} chosen abstains while expert ξ_j predicts and satisfies $\widehat{p}^{t-1}_{I_t,j} \leqslant \gamma_{j,t-1}$, then we do not have access to the true label y_t. In that case, we update optimistically our empirical estimate as if the expert had loss 0 at that round (Step (*) in Alg. 2).

The feedback graph G_t just described can be defined via the out-neighborhood of vertex ξ_j: $N_t(j) = \{\xi_i \in \mathcal{E} \colon \widehat{p}^{t-1}_{j,i} \leqslant \gamma_{i,t-1}\}$. The following regret guarantee holds for UCB-GT.

Theorem 4 *For any $t \in [T]$, let the feedback graph G_t be defined by the out-neighborhood $N_t(j) = \{\xi_i \in \mathcal{E} \colon \widehat{p}^{t-1}_{j,i} \leqslant \gamma_{i,t-1}\}$. Then, the regret of* UCB-GT *is bounded as follows:*

$$\mathcal{O}\left(\mathbb{E}\left[\min_{\mathcal{C} \in \mathcal{F}} \sum_{C \in \mathcal{C}} \frac{\max_{j \in C} \Delta_j}{\min_{j \in C} \Delta_j^2} \log(T) + K\right]\right).$$

Since the graph G_t of UCB-GT has more edges than G^{SUB}, it admits at least as many admissible coverings as G^{SUB}, which leads to a more favorable guarantee than that of UCB-NT run with G^{SUB}. The proof of this result differs from the standard UCB analysis and that of Theorem 3 in that it involves showing that the UCB-GT algorithm can adequately control the amount of bias introduced by the skewed loss estimates. The experiments in the next section provide an empirical validation of this theoretical comparison.

5. Experiments

In this section, we report the results of several experiments on ten datasets comparing UCB-GT, UCB-NT with feedback graph G^{SUB}, vanilla UCB (with no sharing information across experts), as well as Full-Supervision, FS. FS is an algorithm that at each round chooses the expert ξ_j with the smallest abstention loss so far, $\widehat{\mu}_{j,t-1}$, and even if this expert abstains, the algorithm receives the true label and can update the empirical abstention loss estimates for all experts. FS reflects an unrealistic and overly optimistic scenario that clearly falls outside the abstention setting, but it provides an upper bound for the best performance we may hope for.

We used the following eight datasets from the UCI data repository: `HIGGS`, `phishing`, `ijcnn`, `covtype`, `eye`, `skin`, `cod-rna`, and `guide`. We also used the `CIFAR` dataset from (Krizhevsky et al., 2009), where we extracted the first twenty-five principal components and used their projections as features, and a synthetic dataset of points drawn according to the uniform distribution in $[-1, 1]^2$. For

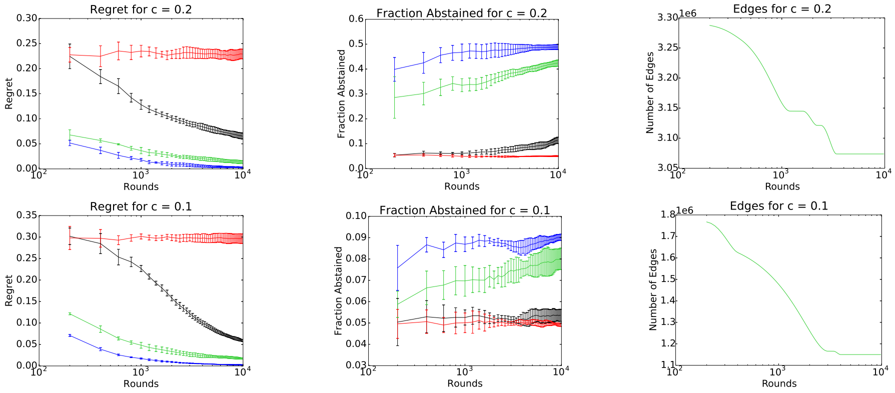

Figure 5: From the left, we show graphs of the average regret $R_t(\cdot)/t$, fraction of points the chosen expert abstained on, and the number of edges of the feedback graph as a function of t (log-scale) for UCB-GT, UCB-NT, UCB, and FS. Top row is the results for `cod-rna` for cost $c = 0.2$ and bottom row is the `guide` for cost $c = 0.1$. More results are in Appendix D.

each dataset, we generated a total of $K = 2{,}100$ experts and all the algorithms were tested for a total of $T = 10{,}000$ rounds. The experts, $\xi = (h, r)$, were chosen in the following way. The predictors h are hyperplanes centered at the origin whose normal vector in $\mathbb{R}^d$ is drawn randomly from the Gaussian distribution, $\mathcal{N}(0, 1)^d$, where d is the dimension of the feature space of the dataset. The abstention functions r are concentric annuli around the origin with radii in $(0, \frac{\sqrt{d}}{20}, \frac{2\sqrt{d}}{20} \ldots, \sqrt{d})$. For each dataset, we generated 100 predictors and each predictor h is paired with the 21 abstention functions r. For a fixed set of experts, we first calculated the regret by averaging over five random draws of the data, where the best-in-class expert was determined in hindsight as the one with the minimum average cumulative abstention loss. We then repeated this experiment five times over different sets of experts and averaged the results. We report these results for $c \in \{0.1, 0.2, 0.3\}$.

Figure 5 shows the averaged regret $R_t(\cdot)/t$ with standard deviations across the five repetitions for the different algorithms as a function of $t \in [T]$ for two datasets. In Appendix D, we present plots of the regret for all ten datasets. These results show that UCB-GT outperforms both UCB-NT and UCB on all datasets for all abstention cost values. Remarkably, UCB-GT's performance is close to that of FS for most datasets, thereby implying that UCB-GT attains almost the best regret that we could hope for. We also find that UCB-NT performs better than the vanilla UCB.

Figure 5 also illustrates the fraction of points in which the chosen expert abstains, as well as the number of edges in the feedback graph as a function of rounds. We only plot the number of edges of UCB-GT since that is the only graph that varies with time. For both experiments depicted and in general for the rest of the datasets, the number of edges for UCB-GT is between 1 million to 3 million, which is at least a factor of 5 more than for UCB-NT, where the number of edges we observed are of the order 200,000. FS enjoys the full information property and the number of edges is fixed at 4 million (complete graph). The increased information sharing of UCB-GT is clearly a strong contributing factor to the algorithm's improvement in regret relative to UCB-NT. In general, we find that, provided that the estimation bias is controlled, the higher is the number of edges, the smaller the regret. Regarding the value of the cost c, as expected, we observe that the fraction of points that the chosen expert abstains on always decreases as c increases, but also that that fraction depends on the dataset and the experts used.

Finally, Appendix D includes more experiments for different aspects of the problem. In particular, we tested how the number of experts or a different choice of experts (confidence-based experts) affected the results. We also experimented with some extreme abstention costs and, as expected, found the fraction of abstained points to be large for $c = 0.001$ and small for $c = 0.9$. In all of these additional experiments, UCB-GT outperformed UCB-NT.

6. Conclusion

We presented a comprehensive analysis of the novel setting of online learning with abstention, including algorithms with favorable guarantees both in the stochastic and adversarial scenarios, and extensive experiments demonstrating the performance of UCB-GT in practice. Our algorithms and analysis can be straightforwardly extended to similar problems, including the multi-class and regression settings, as well as other related scenarios, such as online learning with budget constraints. A key idea behind the design of our algorithms in the stochastic setting is to leverage the stochastic sequence of feedback graphs. This idea can perhaps be generalized and applied to other problems where time-varying feedback graphs naturally appear. Furthermore, our regret guarantees can be instead expressed in terms of the independence number of time-varying graphs by proceeding as in (Lykouris et al., 2019).

References

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. From bandits to experts: A tale of domination and independence. In *NIPS*, 2013.

Alon, N., Cesa-Bianchi, N., Gentile, C., Mannor, S., Mansour, Y., and Shamir, O. Nonstochastic multi-armed bandits with graph-structured feedback. In *CoRR*, 2014.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. Online learning with feedback graphs: Beyond bandits. *JMLR*, 2015.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multi-armed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002b.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

Bartlett, P. and Wegkamp, M. Classification with a reject option using a hinge loss. *JMLR*, pp. 291–307, 2008.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. Leveraging side observations in stochastic bandits. In *UAI*, 2012.

Cesa-Bianchi, N., Gaillard, P., Gentile, C., and Gerchinovitz, S. Algorithmic chaining and the role of partial feedback in online nonparametric learning. In *MLR*, 2017.

Chow, C. An optimum character recognition system using decision function. *IEEE T. C.*, 1957.

Chow, C. On optimum recognition error and reject trade-off. *IEEE T. C.*, 1970.

Clarkson, K. L. Nearest-neighbor searching and metric space dimensions. In *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*. MIT Press, 2006.

Cohen, A., Hazan, T., and Koren, T. Online learning with feedback graphs without the graphs. In *ICML*, 2016.

Cortes, C., DeSalvo, G., and Mohri, M. Learning with rejection. In *ALT*, pp. 67–82. Springer, Heidelberg, Germany, 2016a.

Cortes, C., DeSalvo, G., and Mohri, M. Boosting with abstention. In *NIPS*. MIT Press, 2016b.

El-Yaniv, R. and Wiener, Y. On the foundations of noise-free selective classification. *JMLR*, 2010.

El-Yaniv, R. and Wiener, Y. Agnostic selective classification. In *NIPS*, 2011.

Hazan, E. *Introduction to Online Convex Optimization*. Foundations and Trends in Optimization. Now Publishers Inc., 2016.

Hazan, E. and Megiddo, N. Online learning with prior knowledge. In *COLT*, pp. 499–513, 2007.

Kocák, T., Neu, G., Valko, M., and Munos, R. Efficient learning by implicit exploration in bandit problems with side observations. In *NIPS*, pp. 613–621, 2014.

Krizhevsky, A., Nair, V., and Hinton, G. CIFAR-10 (Canadian Institute for Advanced Research), 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Li, L., Littman, M., and Thomas, W. Knows What It Knows: A framework for self-aware learning. In *ICML*, 2008.

Littlestone, N. and Warmuth, M. K. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Lykouris, T., Tardos, E., and Wali, D. Feedback graph regret bounds for thompson sampling and ucb. In *ArXiv*, 2019.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. In *NIPS*, pp. 291–307, 2011.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *NIPS*, pp. 3168–3176, 2015.

Sayedi, A., Zadimoghaddam, M., and Blum, A. Trading off mistakes and don't-know predictions. In *NIPS*, 2010.

Zhang, C. and Chaudhuri, K. The extended Littlestone's dimension for learning with mistakes and abstentions. In *COLT*, 2016.

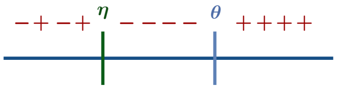

Figure 6: Simple example of the benefits of learning with abstention (Cortes et al., 2016a).

A. Further Related Work

Learning with abstention is a useful paradigm in applications where the cost of misclassifying a point is high. More concretely, suppose the cost of abstention c is less than $1/2$ and consider the set of points along the real line illustrated in Figure 6 where $+$ and $-$ indicate their labels. The best threshold classifier is the hypothesis given by threshold θ, since it correctly classifies points to the right of η, with an expected loss of $(1/2)\,\mathbb{P}[x \leqslant \eta]$. On the other hand, the best abstention pair (h, r) would abstain on the region left of η and correctly classify the rest, with an expected loss of $c\,\mathbb{P}(x \leqslant \eta)$. Since $c < 1/2$, the abstention pair always admits a better loss then the best threshold classifier.

Within the online learning literature, work related to our scenario includes the KWIK (*knows what it knows*) framework of Li et al. (2008) in which the learning algorithm is required to make only correct predictions but admits the option of abstaining from making a prediction. The objective is then to learn a concept exactly with the fewest number of abstentions. If in our framework we received the label at every round, KWIK could be seen as a special case of our framework for online learning with abstention with an infinite misclassification cost and some finite abstention cost. A relaxed version of the KWIK framework was introduced and analyzed by Sayedi et al. (2010) where a fixed number k of incorrect predictions are allowed with a learning algorithm related to the solution of the 'mega-egg game puzzle'. A theoretical analysis of learning in this framework was also recently given by Zhang & Chaudhuri (2016). Our framework does not strictly cover this relaxed framework. However, for some choices of the misclassification cost depending on the horizon, the framework is very close to ours. The analysis in these frameworks was given in terms of mistake bounds since the problem is assumed to be realizable. We will not restrict ourselves to realizable problems and, instead, will provide regret guarantees.

B. Additional material for the adversarial setting

We first present the pseudocode and proofs for the finite arm setting and next analyze the infinite arm setting.

B.1. Finite arm setting

Algorithm 3 contains the pseudocode for EXP3-ABS, an algorithm for online learning with abstention under an adversarial data model that guarantees small regret. The algorithm itself is a simple adaptation of the ideas in (Alon et al., 2014; 2015), where we incorporate the side information that the loss of an abstaining arm is always observed, while the loss of a predicting arm is observed only if the algorithm actually plays a predicting arm. In the pseudocode and in the proof that follows, $L_t(\xi_j)$ is a shorthand for $L(\xi_j, (x_t, y_t))$.

Proof of Theorem 1.
Proof. By applying the standard regret bound of Hedge (e.g., (Bubeck & Cesa-Bianchi, 2012)) to distributions $q_1, \ldots, q_T$ generated by EXP3-ABS and to the non-negative loss estimates $\widehat{L}_t(\xi_j)$, the following holds:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\xi_j\in\mathcal{E}} q_t(\xi_j)\,\mathbb{E}\left[\widehat{L}_t(\xi_j)\right] - \sum_{t=1}^{T}\mathbb{E}\left[\widehat{L}_t(\xi^\star)\right]\right] \leqslant \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{\xi_j\in\mathcal{E}} q_t(\xi_j)\,\mathbb{E}\left[\widehat{L}_t(\xi_j)^2\right]\right], \tag{2}$$

for any fixed $\xi^\star \in \mathcal{E}$. Using the fact that $\mathbb{E}\left[\widehat{L}_t(\xi_j)\right] = L_t(\xi_j)$ and $\mathbb{E}\left[\widehat{L}_t(\xi_j)^2\right] = \frac{L_t(\xi_j)^2}{P_t(\xi_j)}$, we can write

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{\xi_j\in\mathcal{E}} q_t(\xi_j)L_t(\xi_j) - \sum_{t=1}^{T} L_t(\xi^\star)\right] \leqslant \frac{\log K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{\xi_j\in\mathcal{E}} \frac{q_t(\xi_j)}{P_t(\xi_j)} L_t(\xi_j)^2\right].$$

For each t, we can split the nodes V of G_t^{ABS} into the two subsets $V_{abs,t}$ and $V_{acc,t}$ where if a node ξ_j is abstaining at time t

ALGORITHM 3: EXP3-ABS

```
input  Set of experts $\mathcal{E} = \{\xi_1, \ldots, \xi_K\}$; learning rate $\eta > 0$ ;
  Init: $q_1$ is the uniform distribution over $\mathcal{E}$ ;
  for $t \leftarrow 1, 2, \ldots$ do
    RECEIVE($x_t$);
    $\xi_{I_t} \leftarrow$ SAMPLE($q_t$);
    if $r_{I_t}(x_t) > 0$ then
      RECEIVE($y_t$);
    end if
    For all $\xi_j = (h_j, r_j)$, set :
        $P_t(\xi_j) \leftarrow \begin{cases} 1 & \text{if } r_j(x_t) \leqslant 0 \\ \sum_{\xi_i \in \mathcal{E} \,:\, r_i(x_t) > 0} q_t(\xi_i) & \text{if } r_j(x_t) > 0 \,, \end{cases}$
        $\widehat{L}_t(\xi_j) \leftarrow \frac{L_t(\xi_j)}{P_t(\xi_j)} \left( 1_{r_{I_t}(x_t) \leqslant 0} 1_{r_j(x_t) \leqslant 0} + 1_{r_{I_t}(x_t) > 0} \right) \,,$
        $q_{t+1}(\xi_j) \leftarrow \frac{q_t(\xi_j) \exp(-\eta \widehat{L}_t(\xi_j))}{\sum_{\xi_i \in \mathcal{E}} q_t(\xi_i) \exp(-\eta \widehat{L}_t(\xi_i))} \,.$
  end for
```

ALGORITHM 4: CONTEXP3-ABS.

```
input  Ball radius $\varepsilon > 0$, $\varepsilon$-covering $\mathcal{Y}_\varepsilon$ of $\mathcal{Y}$ such that $|\mathcal{Y}_\varepsilon| \leq C_{\mathcal{Y}}\, \varepsilon^{-2}$;
  for $t = 1, 2, \ldots$ do
    RECEIVE($x_t$);
    If $x_t$ does not belong to any existing ball, create new ball of radius $\varepsilon$ centered on $x_t$, and allocate fresh instance of EXP3-ABS;
    Let “Active EXP3-ABS” be the instance allocated to the existing ball whose center $x_s$ is closest to $x_t$;
    Draw action $\xi_{I_t} \in \mathcal{Y}_\varepsilon$ using Active EXP3-ABS;
    Get loss feedback associated with $\xi_{I_t}$ and use it to update state of “Active EXP3-ABS”.
  end for
```

then $\xi_j \in V_{abs,t}$, and otherwise $\xi_j \in V_{acc,t}$. Thus, for any round t, we can write

$$
\begin{aligned}
\sum_{\xi_j \in \mathcal{E}} \frac{q_t(\xi_j)}{P_t(\xi_j)} L_t(\xi_j)^2 &= \sum_{\xi_j \in V_{abs,t}} \frac{q_t(\xi_j)}{P_t(\xi_j)} L_t(\xi_j)^2 + \sum_{\xi_j \in V_{acc,t}} \frac{q_t(\xi_j)}{P_t(\xi_j)} L_t(\xi_j)^2 \\
&\leqslant \sum_{\xi_j \in V_{abs,t}} q_t(\xi_j)\, c^2 + \sum_{\xi_j \in V_{acc,t}} \frac{q_t(\xi_j)}{P_t(\xi_j)} \\
&\leqslant c^2 + 1 \,.
\end{aligned}
$$

The first inequality holds since if ξ_j is an abstaining expert at time t, we know that $L_t(\xi_j) = c$ and $P_t(\xi_j) = 1$, while for the accepting experts we know that $L_t(\xi_j) \leqslant 1$ anyway. The second inequality holds because if ξ_j is an accepting expert, we have $P_t(\xi_j) = \sum_{\xi_j \in V_{acc,t}} q_t(\xi_j)$. Combining this inequality with (2) concludes the proof. □

B.2. Infinite arm setting

Here, the input space $\mathcal{X}$ is assumed to be totally bounded, so that there exists a constant $C_{\mathcal{X}} > 0$ such that, for all $0 < \varepsilon \leqslant 1$, $\mathcal{X}$ can be covered with at most $C_{\mathcal{X}} \varepsilon^{-d}$ balls of radius ε. Let $\mathcal{Y}$ be a shorthand for $[-1, 1]^2$, the range space of the pairs (h, r). An ε-covering $\mathcal{Y}_\varepsilon$ of $\mathcal{Y}$ with respect to the Euclidean distance on $\mathcal{Y}$ has size $K_\varepsilon \leqslant C_{\mathcal{Y}} \varepsilon^{-2}$ for some constant $C_{\mathcal{Y}}$.

The online learning scenario for the loss $\widetilde{L}$ under the abstention setting's feedback graphs is as follows. Given an unknown sequence $z_1, z_2, \ldots$ of pairs $z_t = (x_t, y_t) \in \mathcal{X} \times \{\pm 1\}$, for every round $t = 1, 2, \ldots$:

1. The environment reveals input $x_t \in \mathcal{X}$;
2. The learner selects an action $\xi_{I_t} \in \mathcal{Y}$ and incurs loss $\widetilde{L}(\xi_{I_t}, z_t)$;
3. The learner obtains feedback from the environment.

Our algorithm is described as Algorithm 4. The algorithm essentially works as follows. At each round t, if a new incoming input $x_t \in \mathcal{X}$ is not contained in any existing ball generated so far, then a new ball centered at x_t is created, and a new instance of EXP3-ABS is allocated to handle x_t. Otherwise, the EXP3-ABS instance associated with the closest input so far is used. Each allocated EXP3-ABS instance operates on the discretized action space $\mathcal{Y}_\varepsilon$.

Consider the function

$$\widetilde{L}(a,r) = \begin{cases} c & \text{if } r \leqslant -\gamma \\ 1 + \left(\frac{1-c}{\gamma}\right) r & \text{if } r \in (-\gamma, 0) \\ 1 - \left(\frac{1-f_\gamma(-a)}{\gamma}\right) r & \text{if } r \in [0,\gamma) \\ f_\gamma(-a) & \text{if } r \geqslant \gamma \,, \end{cases}$$

where f_γ is the Lipschitz variant of the 0/1-loss mentioned in Section 3 of the main text (Figure 2 (a)). For any fixed a, the function $\widetilde{L}(a,r)$ is $1/\gamma$-Lipschitz when viewed as a function of r, and is $1/(2\gamma)$-Lipschitz for any fixed r when viewed as a function of a. Hence

$$\begin{aligned} |\widetilde{L}(a,r) - \widetilde{L}(a',r')| &\leqslant |\widetilde{L}(a,r) - \widetilde{L}(a,r')| + |\widetilde{L}(a,r') - \widetilde{L}(a',r')| \\ &\leqslant \frac{1}{\gamma}\,|r-r'| + \frac{1}{2\gamma}\,|a-a'| \\ &\leqslant \sqrt{\frac{1}{\gamma^2} + \frac{1}{4\gamma^2}}\,\sqrt{(a-a')^2 + (r-r')^2} \\ &< \frac{2}{\gamma}\,\sqrt{(a-a')^2 + (r-r')^2}\,, \end{aligned}$$

so that $\widetilde{L}$ is $\frac{2}{\gamma}$-Lipschitz w.r.t. the Euclidean distance on $\mathcal{Y}$. Furthermore, a quick comparison to the abstention loss

$$L(a,r) = f_\gamma(a)1_{r>0} + c1_{r\leqslant 0}$$

reveals that (recall Figure 2 (b) in the main text) :

- $\widetilde{L}$ is an upper bound on L, i.e.,
$$\widetilde{L}(a,r) \geqslant L(a,r), \quad \forall\,(a,r) \in \mathcal{Y}\,;$$
- $\widetilde{L}$ approximates L in that
$$\widetilde{L}(a,r) = L(a,r), \quad \forall\,(a,r) \in \mathcal{Y} \,:\, |r| \geqslant \gamma\,. \tag{3}$$

With the above properties of $\widetilde{L}$ at hand, we are ready to prove Theorem 2.

Proof of Theorem 2.
Proof. On each ball $B \subseteq \mathcal{X}$ that CONTEXP3-ABS allocates during its online execution, Theorem 1 supplies the following regret guarantee for the associated instance of EXP3-ABS:

$$\frac{\log K_\varepsilon}{\eta} + \frac{\eta}{2} T_B(c^2+1)\,,$$

where T_B is the number of points x_t falling into ball B. Now, taking into account that $\widetilde{L}$ is $\frac{2}{\gamma}$-Lipschitz, and that the functions h and r are assumed to be $L_\mathcal{E}$-Lipschitz on $\mathcal{X}$, a direct adaptation of the proof of Theorem 1 in (Cesa-Bianchi et al., 2017) gives the bound

$$\sup_{\xi \in \mathcal{E}} \mathbb{E}\left[\sum_{t=1}^{T} \widetilde{L}(\xi_{I_t}, z_t) - \sum_{t=1}^{T} \widetilde{L}(\xi, z_t)\right] \leqslant \frac{N_T \log K_\varepsilon}{\eta} + \frac{\eta}{2} T(c^2+1) + L_\mathcal{E}\,\varepsilon\,\frac{2}{\gamma}\,T\,,$$

being $N_T \leqslant C_{\mathcal{X}}\varepsilon^{-d}$ the maximum number of balls created by CONTEXP3-ABS. Using $c \leqslant 1$ and setting $\eta = \sqrt{\frac{N_T\,\log K_\varepsilon}{T}}$ yields

$$\sup_{\xi\in\mathcal{E}} \mathbb{E}\left[\sum_{t=1}^T \widetilde{L}(\xi_{I_t}, z_t) - \sum_{t=1}^T \widetilde{L}(\xi, z_t)\right] \leqslant 2\sqrt{T\,N_T\,\log K_\varepsilon} + L_{\mathcal{E}}\,\varepsilon\,\frac{2}{\gamma}\,T\,.$$

Next, optimizing for ε by setting $\varepsilon \simeq T^{-\frac{1}{2+d}}\left(\frac{1}{\gamma}\right)^{-\frac{2}{2+d}}$ (and disregarding $L_{\mathcal{E}}$ and log factors) gives

$$\sup_{\xi\in\mathcal{E}} \mathbb{E}\left[\sum_{t=1}^T \widetilde{L}(\xi_{I_t}, z_t) - \sum_{t=1}^T \widetilde{L}(\xi, z_t)\right] = \widetilde{\mathcal{O}}\left(T^{\frac{d+1}{d+2}}\left(\frac{1}{\gamma}\right)^{\frac{d}{d+2}}\right)\,. \tag{4}$$

Finally, we are left with connecting the above bound on the regret with a bound on the regret for L. Now, observe that

$$\mathbb{E}\left[\sum_{t=1}^T \widetilde{L}(\xi_{I_t}, z_t)\right] \geqslant \mathbb{E}\left[\sum_{t=1}^T L(\xi_{I_t}, z_t)\right]\,, \tag{5}$$

since $\widetilde{L}(\xi, z_t)$ is an upper bound on $L(\xi, z_t)$ for any ξ and z_t. Moreover, if we assume for the sake of brevity that the minima are reached (the general case is straightforward to handle in a similar way), we can define

$$\xi^* = (h^*, r^*) = \operatorname*{argmin}_{\xi\in\mathcal{E}} \sum_{t=1}^T L(\xi, z_t), \qquad \widetilde{\xi}^* = \operatorname*{argmin}_{\xi\in\mathcal{E}} \sum_{t=1}^T \widetilde{L}(\xi, z_t)\,.$$

We denote by $M_T^*(\gamma)$ the number of x_t such that $|r^*(x_t)| \leqslant \gamma$. Then, we can write

$$\begin{aligned}
\sum_{t=1}^T \widetilde{L}(\widetilde{\xi}^*, z_t) &\leqslant \sum_{t=1}^T \widetilde{L}(\xi^*, z_t)\\
&\leqslant \sum_{t\,:\,|r^*(x_t)|>\gamma}^T \widetilde{L}(\xi^*, z_t) + M_T^*(\gamma)\\
&\quad \text{(since } \widetilde{L} \leqslant 1\text{))}\\
&= \sum_{t\,:\,|r^*(x_t)|>\gamma}^T L(\xi^*, z_t) + M_T^*(\gamma)\\
&\quad \text{(using (3))}\\
&\leqslant \sum_{t=1}^T L(\xi^*, z_t) + M_T^*(\gamma)\,.
\end{aligned}$$

Combining with (4) and (5) gives the following regret bound

$$\sup_{\xi\in\mathcal{E}} \mathbb{E}\left[\sum_{t=1}^T L(\xi_{I_t}, z_t) - \sum_{t=1}^T L(\xi, z_t)\right] \leqslant \widetilde{\mathcal{O}}\left(T^{\frac{d+1}{d+2}}\left(\frac{1}{\gamma}\right)^{\frac{d}{d+2}}\right) + M_T^*(\gamma)\,,$$

thereby concluding the proof. □

Remark 1 *The reader should observe that, since the algorithm is competing against an* uncountably infinite *set of experts, the standard regret guarantee of $\sqrt{T}$ that one can achieve in the finite case cannot be obtained in general (see, e.g., the lower bound on regret of $T^{(d-1)/d}$ by (Hazan & Megiddo, 2007), which holds in the easier full information setting). Notice that, while our algorithm* CONTEXP3-ABS *admits a slightly worse bound of the form $T^{(d+1)/(d+2)}$, it has the advantage of being computationally feasible. In particular, the covering of the input space $\mathcal{X}$ can be done adaptively, as the points x_t are observed. In doing so, the number of ε-balls allocated can never exceed the total number of rounds T. Given a new x_t, the algorithm has to decide if a new ball needs to be created or an old ball can be used. Known data-structures exist to efficiently implement this decision (e.g., (Clarkson, 2006)). The extra additive term $M_T^*(\gamma)$ in Theorem 2 is due to the* *fact that the loss function L therein is not Lipschitz. In fact, one can further improve the term $T^{\frac{d+1}{d+2}}$ to $T^{\frac{d}{d+1}}$ by adopting a hierarchical covering technique of the function space $\mathcal{E}$, each layer of the hierarchy being a pool of experts for the layer above it, see, e.g., (Cesa-Bianchi et al., 2017). However, the resulting algorithm would be of theoretical interest only, since it would be computationally very costly.*

C. Additional material for the stochastic setting

In this section, we present the proofs of the theoretical guarantees for UCB-NT and UCB-GT, as well as the proof of Proposition 1. The following theorems hold more generally with $S_{j,t} = \sqrt{\frac{2\beta \log t}{Q_{j,t}}}$ for $\beta > 2$, which implies slightly better constants in the regret bound. However, for the sake of the simplicity of the presentation, below we set $\beta = \frac{5}{2}$. Moreover, we prove Theorem 3 for the abstention loss L, but it holds for any general loss function.

C.1. Regret of UCB-NT

Proof of Theorem 3.

Proof. Consider a sequence of graph realizations $G_1, \ldots, G_t$ denoted by $\mathbf{G}_t$. By conditioning on this quantity, the regret can be decomposed according to each arm i:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[L(\xi_{I_t}, z_t) - L(\xi_*, z_t)] &= \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[L(\xi_{I_t}, z_t) - L(\xi_*, z_t)|\mathbf{G}_t]] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{K} 1_{I_t=i}(L(\xi_i, z_t) - L(\xi_*, z_t))\middle|\mathbf{G}_t\right]\right] \\
&= \sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[L(\xi_i, z_t) - L(\xi_*, z_t)|\mathbf{G}_t]\,\mathbb{E}[1_{I_t=i}|\mathbf{G}_t]] \\
&= \sum_{i=1}^{K}\sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[L(\xi_i, z_t) - L(\xi_*, z_t)]\,\mathbb{E}[1_{I_t=i}|\mathbf{G}_t]] = \mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{T} \Delta_i\,\mathbb{E}[1_{I_t=i}|\mathbf{G}_t]\right]
\end{aligned}$$

where, in the last step, we used the fact that $L(\cdot, z_t)$s are independent of $\mathbf{G}_t$ since, by assumption, $\mathbf{G}_t$ only depends on information up $t-1$. Next, we focus on bounding $\sum_{t=1}^{T} \mathbb{E}[1_{I_t=i}|\mathbf{G}_t]$ for each arm i.

We split the expectation according to the events $Q_{i,t-1} > s_i$ and $Q_{i,t-1} \leqslant s_i$, where s_i is a quantity determined later:

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}[1_{I_t=i}|\mathbf{G}_t] &= \sum_{t=1}^{T} \mathbb{E}[1_{I_t=i}(1_{Q_{i,t-1}\leqslant s_i} + 1_{Q_{i,t-1}>s_i})|\mathbf{G}_t] \\
&\leqslant s_i + \sum_{t=1}^{T} \mathbb{E}[1_{I_t=i}1_{Q_{i,t-1}>s_i}|\mathbf{G}_t].
\end{aligned}$$

We wish to choose s_i sufficiently large so that the second term is bounded but so that it admits a mild dependence on T. Now, whenever $I_t = i$, by the design of the algorithm, it must be the case that the upper confidence bound of i is smaller than that of any other expert. Thus,

$$\mathbb{E}[1_{I_t=i}1_{Q_{i,t-1}>s_i}|\mathbf{G}_t] = \mathbb{P}[I_t = i, Q_{i,t-1} > s_i|\mathbf{G}_t] \leqslant \mathbb{P}[\widehat{\mu}_{i,t-1} - S_{i,t-1} \leqslant \widehat{\mu}_{*,t-1} - S_{*,t-1}, Q_{i,t-1} > s_i|\mathbf{G}_t],$$

where $*$ denotes the best-in-class expert. We now use the terms μ_*, μ_i and $S_{i,t-1}$ to reorder the first event in the probability on the right-hand side of the last expression as follows:

$$\begin{aligned}
&0 \leqslant \widehat{\mu}_{*,t-1} - S_{*,t-1} - \widehat{\mu}_{i,t-1} + S_{i,t-1} \\
\Leftrightarrow\ &0 \leqslant (\widehat{\mu}_{*,t-1} - S_{*,t-1} - \mu_*) + (\mu_i - \widehat{\mu}_{i,t-1} + S_{i,t-1} - 2S_{i,t-1}) + (\mu_* - \mu_i + 2S_{i,t-1})\,.
\end{aligned}$$

If we can show that the third term is negative, then the first and second term must be positive. Moreover, we will further show that the first and second terms can only be positive with an extremely low probability that is bounded by a constant independent of T. Furthermore, the third term will be negative whenever the slack term in the upper confidence bound is small enough, which amounts to choosing s_i large enough.

In particular, by setting $s_i = \frac{20\log(T)}{\Delta_i^2}$, we ensure that the event $Q_{i,t-1} > s_i$ implies that

$$Q_{i,t-1} > \frac{20\log(t)}{\Delta_i^2} \Leftrightarrow \mu_* - \mu_i + 2S_{i,t-1} < 0.$$

As explained above, it then follows that

$$\begin{aligned}
&\mathbb{P}[\widehat{\mu}_{i,t-1} - S_{i,t-1} \leqslant \widehat{\mu}_{*,t-1} - S_{*,t-1}, Q_{i,t-1} > s_i|\mathbf{G}_t] \\
&\leqslant \mathbb{P}[\widehat{\mu}_{*,t-1} - S_{*,t-1} - \mu_* \geqslant 0|\mathbf{G}_t] + \mathbb{P}[\mu_i - \widehat{\mu}_{i,t-1} + S_{i,t-1} - 2S_{i,t-1} \geqslant 0|\mathbf{G}_t].
\end{aligned}$$

We can bound these last probabilities using the union bound and a concentration inequality such as Hoeffding's Inequality:

$$\begin{aligned}
&\mathbb{P}[\mu_i - \widehat{\mu}_{i,t-1} + S_{i,t-1} - 2S_{i,t-1} \geqslant 0|\mathbf{G}_t] \\
&= \mathbb{P}\Big[- \tfrac{1}{Q_{i,t-1}} \sum_{s=1}^{t-1} L(\xi_i, z_s) 1_{i \in N_s(I_s)} + \mu_i - \sqrt{\tfrac{5\log(t)}{Q_{i,t-1}}} \geqslant 0 \Big| \mathbf{G}_t \Big].
\end{aligned}$$

Now, the estimate $\widehat{\mu}_{i,t-1}$ is an average of i.i.d. realizations of the random variable $L(\xi_i, z)$, with $z \sim \mathcal{D}$, since the out-neighborhood of the chosen expert only depends on previous observations. That is,

$$\begin{aligned}
\frac{\mathbb{E}[\sum_{s=1}^{t-1} L(\xi_i, z_s) 1_{i \in N_s(I_s)}]}{\mathbb{E}[\sum_{s=1}^{t-1} 1_{i \in N_s(I_s)}]} &= \frac{\mathbb{E}[\sum_{s=1}^{t-1} \mathbb{E}[L(\xi_i, z_s) 1_{i \in N_s(I_s)} | i \in N_s(I_s)]]}{\mathbb{E}[\sum_{s=1}^{t-1} 1_{i \in N_s(I_s)}]} \\
&= \frac{\mathbb{E}[\sum_{s=1}^{t-1} 1_{i \in N_s(I_s)} \mathbb{E}[L(\xi_i, z_s) | i \in N_s(I_s)]]}{\mathbb{E}[\sum_{s=1}^{t-1} 1_{i \in N_s(I_s)}]} \\
&= \frac{\mathbb{E}[\sum_{s=1}^{t-1} 1_{i \in N_s(I_s)} \mathbb{E}[L(\xi_i, z_s)]]}{\mathbb{E}[\sum_{s=1}^{t-1} 1_{i \in N_s(I_s)}]} \\
&= \mathbb{E}[L(\xi_i, z)].
\end{aligned}$$

Hence, $\widehat{\mu}_{i,t-1}$ can be turned into an empirical estimate of μ_i using the union bound as follows:

$$\begin{aligned}
\mathbb{P}\Big[- \tfrac{1}{Q_{i,t-1}} \sum_{s=1}^{t-1} L(\xi_i, z_s) 1_{i \in N_s(I_s)} + \mu_i - \sqrt{\tfrac{5\log(t)}{Q_{i,t-1}}} \geqslant 0 \Big| \mathbf{G}_t \Big] &\leqslant \mathbb{P}\Big[\exists n \in [1, t] : -\widehat{\mu}_i^n + \mu_i - \sqrt{\tfrac{5\log(t)}{n}} \Big| \mathbf{G}_t \Big] \\
&\leqslant \sum_{n=1}^{t} \frac{1}{t^{\frac{5}{2}}} = \frac{1}{t^{\frac{3}{2}}} \ ,
\end{aligned}$$

where $\widehat{\mu}_i^n = \frac{1}{n} \sum_{s=1}^{n} L(\xi_i, z_s)$. By the same reasoning, we can also bound the probability of the best arm :

$$\mathbb{P}\Big[\widehat{\mu}_{*,t-1} - S_{*,t-1} - \mu_* \geqslant 0 \Big| \mathbf{G}_t \Big] \leqslant \sum_{n=1}^{t} \frac{1}{t^{\frac{5}{2}}} = \frac{1}{t^{\frac{3}{2}}} \ .$$

Now let $\mathcal{C}$ be any element of $\mathcal{F}$ and C any element of $\mathcal{C}$. Then for any $i \in C$, it follows that

$$Q_{i,t} = \sum_{s=1}^{t} 1_{i \in N_s(I_s)} = \sum_{s=1}^{t} \sum_{j=1}^{K} 1_{i \in N_s(j)} 1_{I_s = j} \geqslant \sum_{s=1}^{t} \sum_{j \in C} 1_{i \in N_s(j)} 1_{I_s = j} = \sum_{s=1}^{t} \sum_{j \in C} 1_{I_s = j}.$$

This implies that

$$\begin{aligned}
&\sum_{i=1}^{K}\sum_{t=1}^{T}\Delta_i\,\mathbb{E}[1_{I_t=i}1_{Q_{i,t-1}\leqslant s_i}|\mathbf{G}_t]\\
&\leqslant \sum_{C\in\mathcal{C}}\left[\max_{i\in C}\Delta_i\right]\sum_{t=1}^{T}\sum_{i\in C}\mathbb{E}\left[1_{I_t=i}1_{Q_{i,t-1}\leqslant s_i}\middle|\mathbf{G}_t\right]\\
&\leqslant \sum_{C\in\mathcal{C}}\left[\max_{i\in C}\Delta_i\right]\sum_{t=1}^{T}\sum_{i\in C}\mathbb{E}\left[1_{I_t=i}1_{Q_{i,t-1}\leqslant \max_{j\in C} s_j}\middle|\mathbf{G}_t\right]\\
&\leqslant \sum_{C\in\mathcal{C}}\left[\max_{i\in C}\Delta_i\right]\max_{j\in C} s_j.
\end{aligned}$$

Combining the above calculations, applying our definition for s_i, and using the fact that the above analysis holds for any shared admissible covering shows that

$$\mathbb{E}\left[\min_{\mathcal{C}\in\mathcal{F}}\sum_{C\in\mathcal{C}}\left[\max_{j\in C}\Delta_j\right]\left[\max_{j\in C}\frac{20\log(T)}{\Delta_j^2}\right]+5K\right],$$

which proves the bound of the theorem. □

C.2. Regret of UCB-GT

Next, we prove the regret bound for UCB-GT, which demonstrates how one can exploit the bias and feedback structure in the problem.

Proof of Theorem 4.
Proof. As in the previous proof, we focus on bounding $\sum_{t=1}^{T}\mathbb{E}[1_{I_t=i}|\mathbf{G}_t]$ for each arm i. We again split the expectation according to the events based on $Q_{i,t-1}$ as follows:

$$\sum_{t=1}^{T}\mathbb{E}[1_{I_t=i}1_{Q_{i,t-1}\leqslant s_i}|\mathbf{G}_t]+\mathbb{E}[1_{I_t=i}1_{Q_{i,t-1}>s_i}|\mathbf{G}_t],$$

where s_i is to be determined later. We then bound the second term using the algorithm's choice of arm, I_t:

$$\mathbb{E}[1_{I_t=i}1_{Q_{i,t-1}>s_i}|\mathbf{G}_t]=\mathbb{P}[I_t=i,Q_{i,t-1}>s_i|\mathbf{G}_t]\leqslant\mathbb{P}[\widehat{\mu}_{i,t-1}-S_{i,t-1}\leqslant\widehat{\mu}_{*,t-1}-S_{*,t-1},Q_{i,t-1}>s_i|\mathbf{G}_t].$$

$\widehat{\mu}_{i,t-1}$ is a biased estimate of μ_i. This is because whenever x_s falls in the region $\{x\colon r_i(x)>0\wedge r_{I_s}(x)\leqslant 0\}$ and the condition $\widehat{p}_{I_s,i}^{s-1}\leqslant\gamma_{i,s-1}$ holds, the label y_s is not accessible. In this case, the UCB-GT algorithm updates the average loss of expert i optimistically, as if the expert were correct at that time step.

We can decompose this biased estimate $\widehat{\mu}_{i,t-1}$ into two terms: $\widehat{\mu}_{i,t-1}=\widetilde{\mu}_{i,t-1}-\varepsilon_{i,t-1}$. The first term, $\widetilde{\mu}_{i,t-1}$, is an unbiased estimate of arm i and similar to the estimates in Theorem 3. The second term is the misclassification rate $\varepsilon_{i,t-1}$ over $\{s\in[t-1]\colon r_i(x_s)>0\cap r_{I_s}(x_s)\leqslant 0\}$ whenever the condition $\widehat{p}_{I_s,i}^{s-1}\leqslant\gamma_{i,s-1}$ holds, that is, $\varepsilon_{i,t-1}=\frac{1}{Q_{i,t-1}}\sum_{s=1}^{t-1}1_{y_sh_i(x_s)\leqslant 0}1_{r_i(x_s)>0,r_{I_s}(x_s)\leqslant 0}1_{\widehat{p}_{I_s,i}^{s-1}\leqslant\gamma_{i,s-1}}$.

Now, by the design of the UCB-GT, if arm i is chosen at time t, it must be the case that $\widehat{\mu}_{i,t-1}-S_{i,t-1}\leqslant\widehat{\mu}_{*,t-1}-S_{*,t-1}$. We can expand and rewrite this expression as follows:

$$\begin{aligned}
&0\leqslant\widehat{\mu}_{*,t-1}+\varepsilon_{i^*,t-1}-\varepsilon_{i^*,t-1}-S_{*,t-1}-\widehat{\mu}_{i,t-1}-\varepsilon_{i,t-1}+\varepsilon_{i,t-1}+S_{i,t-1}\\
\Leftrightarrow\;&0\leqslant(\widetilde{\mu}_{*,t-1}-S_{*,t-1}-\mu_*)+(\mu_i-\widetilde{\mu}_{i,t-1}+S_{i,t-1}-2S_{i,t-1})+(\mu_*-\mu_i+(2+C)S_{i,t-1})\,,
\end{aligned}$$

where we used the fact that $-\varepsilon_{i^*,t-1} \leqslant 0$, and where we bounded $\varepsilon_{i,t-1}$ as follows:

$$\begin{aligned}
\varepsilon_{i,t-1} &= \frac{1}{Q_{i,t-1}} \sum_{s=1}^{t-1} 1_{y_s h_i(x) \leqslant 0} 1_{r_i(x_s)>0, r_{I_s}(x_s) \leqslant 0} 1_{\widehat{p}_{I_s,i}^{s-1} \leqslant \gamma_{i,s-1}} \\
&\leqslant \frac{1}{Q_{i,t-1}} \sum_{s=1}^{t-1} 1_{r_i(x_s)>0, r_{I_s}(x_s) \leqslant 0} 1_{\widehat{p}_{I_s,i}^{s-1} \leqslant \gamma_{i,s-1}} = \frac{1}{Q_{i,t-1}} \sum_{s=1}^{t-1} \sum_{\xi_j \in \mathcal{E}-\xi_i} 1_{r_i(x_s)>0, r_j(x_s) \leqslant 0} 1_{\widehat{p}_{I_s,i}^{s-1} \leqslant \gamma_{i,s-1}} 1_{I_s=j} \\
&\leqslant \frac{1}{Q_{i,t-1}} \sum_{\xi_j \in \mathcal{E}-\xi_i} \sum_{s=1}^{t-1} 1_{r_i(x_s)>0, r_j(x_s) \leqslant 0} 1_{\widehat{p}_{I_s,i}^{s-1} \leqslant \gamma_{i,s-1}}.
\end{aligned}$$

The condition $\widehat{p}_{j,i}^{s-1} \leqslant \gamma_{i,s-1}$ is equivalent to $\sum_{k=1}^{s-1} 1_{r_i(x_k)>0, r_j(x_k) \leqslant 0} \leqslant (s-1)\gamma_{i,s-1}$. Since the sum above is non-zero only when this condition holds, there exists $s_j \in [1, t-1]$ such that $\sum_{s=1}^{t-1} 1_{r_i(x_s)>0, r_j(x_s) \leqslant 0} 1_{\widehat{p}_{j,i}^{s} \leqslant \gamma_{i,s}} \leqslant (s_j-1)\gamma_{i,s_j-1} + 1$. Moreover, using the fact that $(s_j-1)\gamma_{i,s_j-1} = \sqrt{5Q_{i,s_j-1}\log(s_j)}/(K-1) \leqslant \sqrt{5Q_{i,t-1}\log(t-1)}/(K-1)$, we can conclude that

$$\varepsilon_{i,t-1} \leqslant \frac{1}{Q_{i,t-1}} \sum_{\xi_j \in \mathcal{E}-\xi_i} \sum_{s=1}^{t-1} 1_{r_i(x_s)>0, r_j(x_s) \leqslant 0} 1_{\widehat{p}_{j,i}^{s-1} \leqslant \gamma_{i,s-1}} \leqslant \frac{K-1}{Q_{i,t-1}} \left[\frac{\sqrt{5Q_{i,t-1}\log(t-1)}}{K-1} + 1 \right] \leqslant C\sqrt{\frac{5\log(t-1)}{Q_{i,t-1}}}$$

for some constant $C > 0$. The rest of the proof now follows by similar arguments as in the proof of Theorem 3. Specifically, we can choose s_i such that the term $\mu_* - \mu_i + (2+C)S_{i,t-1}$ is negative, and since now $\widetilde{\mu}_{*,t-1}$ and $\widetilde{\mu}_{i,t-1}$ are unbiased estimates, we can bound the probabilities $\mathbb{P}[\widetilde{\mu}_{*,t-1} - S_{*,t-1} - \mu_* \geqslant 0 | \mathbf{G}_t]$ and $\mathbb{P}[\mu_i - \widetilde{\mu}_{i,t-1} - S_{i,t-1} \geqslant 0 | \mathbf{G}_t]$ using standard concentration inequalities. □

C.3. Linear regret without the subset property

In this section, we prove Proposition 1, which shows that when the subset property does not hold for a feedback graph, then it is possible to incur linear regret.

Proof of Proposition 1.
Proof. Let $p^* \in (0,1)$. We design a setting in which with probability at least p^*, the UCB-NT algorithm incurs linear regret.

Since the family of abstention functions induces a feedback graph that violates the subset property, there exist pairs (h_i, r_i) and (h_j, r_j) and points x^*, $\widetilde{x}$ for which $x^* \in \mathcal{A}_i \setminus \mathcal{A}_j$, $\widetilde{x} \in \mathcal{A}_i \cap \mathcal{A}_j$, where $\mathcal{A}_i$ and $\mathcal{A}_j$ are the acceptance regions associated with r_i and r_j, respectively, and the feedback graph is designed such that the algorithm updates the pair (h_i, r_i) when the pair (h_j, r_j) is selected.

Now, for some $p \in (0,1)$ to be determined later, consider a distribution with probability p on $(\widetilde{x}, \widetilde{y})$ and $(1-p)$ on (x^*, y^*).

We choose the set of hypothesis functions $\mathcal{H} = \{h_i, h_j\}$, the loss function ℓ in (1), and the labels y^* and $\widetilde{y}$ in such a way that $\ell(\widetilde{y}, h_i(\widetilde{x})) = c - \beta$, $\ell(\widetilde{y}, h_j(\widetilde{x})) = c - \alpha$, and $\ell(y^*, h_i(x^*)) = 0$, where α, β are values that will be later specified. For instance, we can consider the hinge loss $\ell(y, \widehat{y}) = (1 - y\widehat{y})_+$, and h_i, h_j such that $h_i(\widetilde{x}) = \frac{1-c+\beta}{\widetilde{y}}$, $h_j(\widetilde{x}) = \frac{1-c+\alpha}{\widetilde{y}}$, and $h_i(x^*) = \frac{1}{y^*}$. Note that, since $r_j(x^*) < 0$, $\ell(y^*, h_j(x^*))$ may admit any value.

Now, by construction, $\mu_i = (c-\beta)p$ and $\mu_j = (c-\alpha)p + c(1-p) = c - \alpha p$. We claim that we can choose α, β and p such that (1) $\alpha > \beta$; (2) $\mu_i < \mu_j$; (3) $\mu_j < \ell(\widetilde{y}, h_i(\widetilde{x}))$.

The first condition is immediate. The second condition is equivalent to $cp - \beta p < c - \alpha p$, which is itself equivalent to $\alpha - \beta < \frac{c(1-p)}{p}$. By continuity, we can choose α and β close enough such that this is true for any $p \in (0,1)$. The third condition is equivalent to $c - \alpha p < c - \beta$, which is itself equivalent to $\beta < \alpha p$. This is true for p close enough to 1.

Now let $n \in \mathbb{N}$ be large enough such that $\mu_j < \ell(\widetilde{y}, h_i(\widetilde{x})) - \sqrt{\frac{5\log(n)}{n}}$. By continuity, we can choose p large enough such that $p > (p^*)^{1/n}$, and for this choice of p, we can choose α and β such that $\alpha > \beta$, $\alpha, \beta < c$, $\alpha - \beta < \frac{c(1-p)}{p}$, and $\beta < \alpha p$. For instance, if we, without loss of generality, assume that $p > \frac{1}{2}$, then we can choose, $\alpha = \frac{c(1-p)}{2p}$ and $\beta = \frac{c(1-p)}{4}$.

Then, with probability $p^n > p^*$, the point $\widetilde{x}$ will be sampled n times at the start of the game, such that the pair (h_j, r_j) will

Dataset	Number of features
covtype	54
ijcnn	22
skin	3
HIGGS	28
guide	4
phishing	68
cod	8
eye	14
CIFAR	25

Table 1: Table shows the number of features of each dataset.

have a lower confidence bound than the pair (h_i, r_i) at all time steps. Thus, UCB-NT will choose the pair (h_j, r_j) throughout the entire game, even though $\mu_i < \mu_j$. Consequently, the regret of the algorithm will be at least $T(\mu_j - \mu_i)$. □

D. Additional experimental results

In this section, we present several figures showing our experimental results. Figure 7 and Figure 8 show the regret for different abstention costs $c \in \{0.1, 0.2, 0.3\}$ for all our datasets. We observe that, in general, UCB-GT outperforms UCB-NT and UCB for all datasets and is even within the standard deviation of the FS's regret for some of datasets. The figures also indicate that the regret of UCB decreases slowly. This is expected, since there are 2,100 experts, 10,000 time steps, and the algorithm only updates a single expert per time step.

Figure 9 and Figure 10 show the fraction of abstained points for all the datasets. Figure 11 also shows how the fraction of abstained points varies with abstention cost for two extreme values $c \in \{0.001, 0.9\}$. Again UCB-GT admits a lower regret than UCB-NT and UCB and, as expected, the fraction of points decreases as the cost of abstention increases. Figure 12 shows the effect of using confidence-based experts and suggests that the choice of experts does not affect the relative performance of the algorithms. We also tested the effect of varying the number of experts: Figure 13 shows the regret of three datasets when the number of experts is $K = 500$ and $T = 5{,}000$. For this set of experts, we find a similar pattern of performance as above.

Next, we describe in more detail the datasets and how they were processed. In Table 1, we show the number of features for each dataset. For all datasets, we normalized the features to be in the interval $[-1, 1]$. Note that the reason for choosing abstention functions with radius range $(0, \sqrt{d})$ is to cover the entire hypercube $[-1, 1]^d$ with our concentric annuli. For the CIFAR dataset, we extracted the first twenty-five principal components of the horse and boat images, projected the images on these components, and normalized the range of the projections to $[-1, 1]$. The features of the synthetic dataset are drawn from the uniform distribution over $[-1, 1]^2$ and the label is determined by the sign of the projection of a point onto the normal of the diagonal hyperplane $y = -x$.

The confidence-based abstention function has the form $r(x) = |h(x)| - \theta$. In our experiments (Figure 12), we generated twenty abstention functions with thresholds $\theta \in (0, \ldots, 0.25)$, which are paired with each predictor. The predictors are axis-aligned planes along each feature of the dataset. For each dataset, the number of predictors is $\lfloor 100/d \rfloor$ where d is the dimension of the dataset. We chose twenty abstention functions and about 100 prediction functions in order to match the experimental setup of the randomly drawn experts. The total number of experts is then $\lfloor 100/d \rfloor \cdot 20 \cdot d$. Note that we only tested some of our datasets since for larger dimensions d, the number of experts per feature was too small.

D.1. Average regret for different abstention costs and datasets

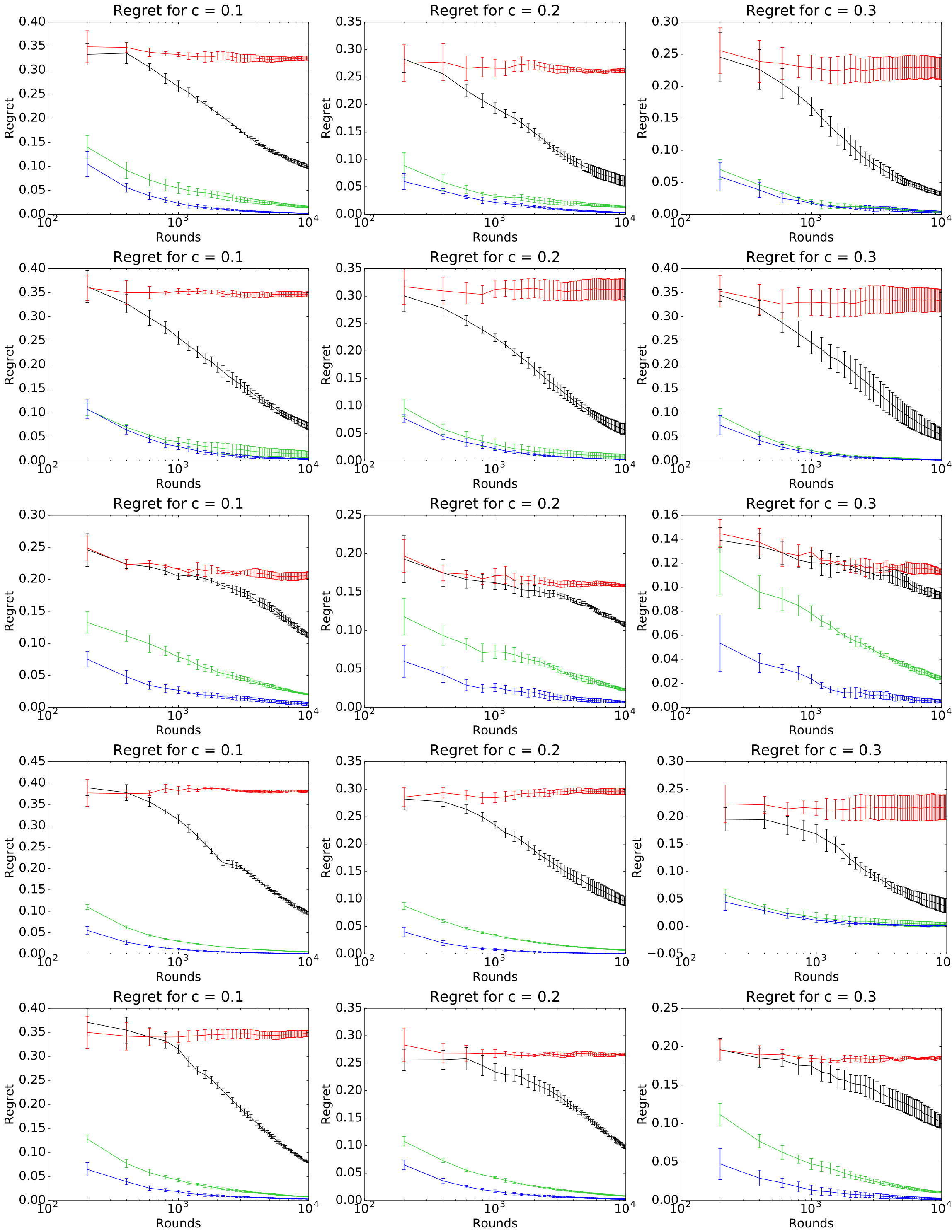

Figure 7: A graph of the averaged regret $R_t(\cdot)/t$ with standard deviations as a function of t (log scale) for UCB-GT, UCB-NT, UCB, and FS for different values of abstention costs. Each row is a dataset, starting from the top row we have: CIFAR, ijcnn, HIGGS, phishing, and covtype.

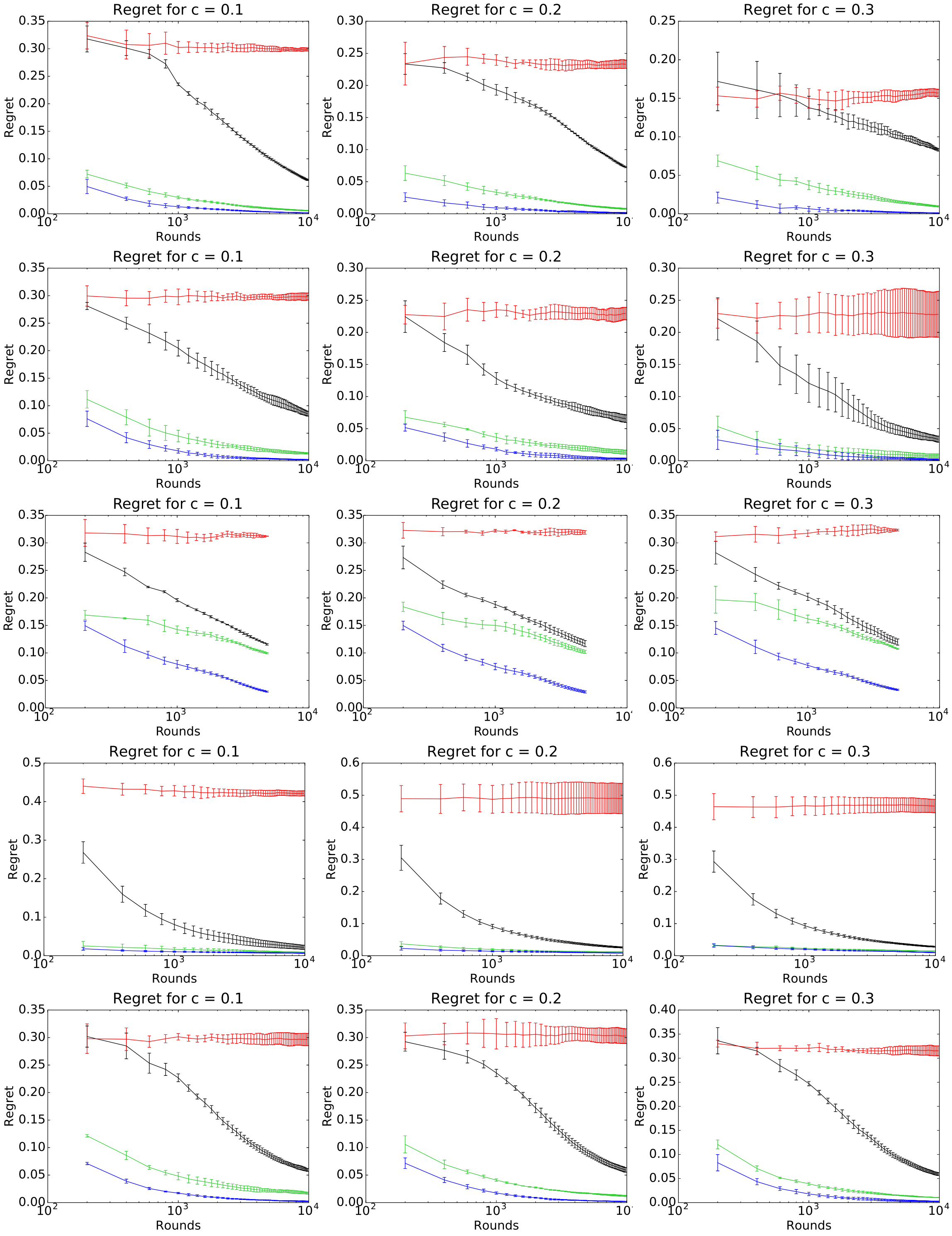

Figure 8: A graph of the averaged regret $R_t(\cdot)/t$ with standard deviations as a function of t (log scale) for UCB-GT, UCB-NT, UCB, and FS for different values of abstention costs. Each row is a dataset, starting from the top row we have: eye, cod-ran, synthetic, skin, and guide.

D.2. Average fraction of abstention points for different abstention costs and datasets

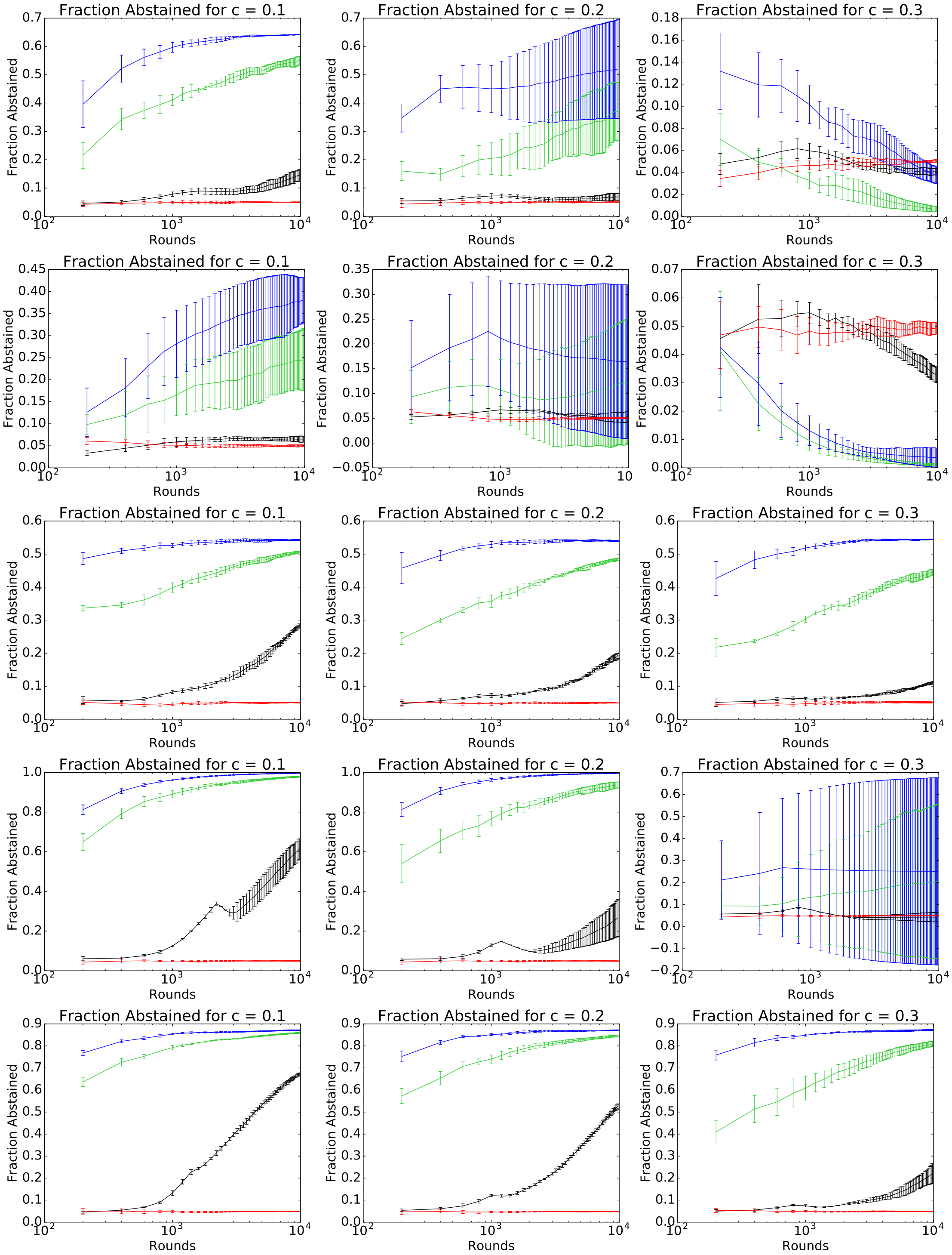

Figure 9: A graph of the averaged fraction of abstained points with standard deviations as a function of t (log scale) for UCB-GT, UCB-NT, UCB, and FS for different values of abstention costs. Each row is a dataset, starting from the top row we have: CIFAR, ijcnn, HIGGS, phishing, and covtype.

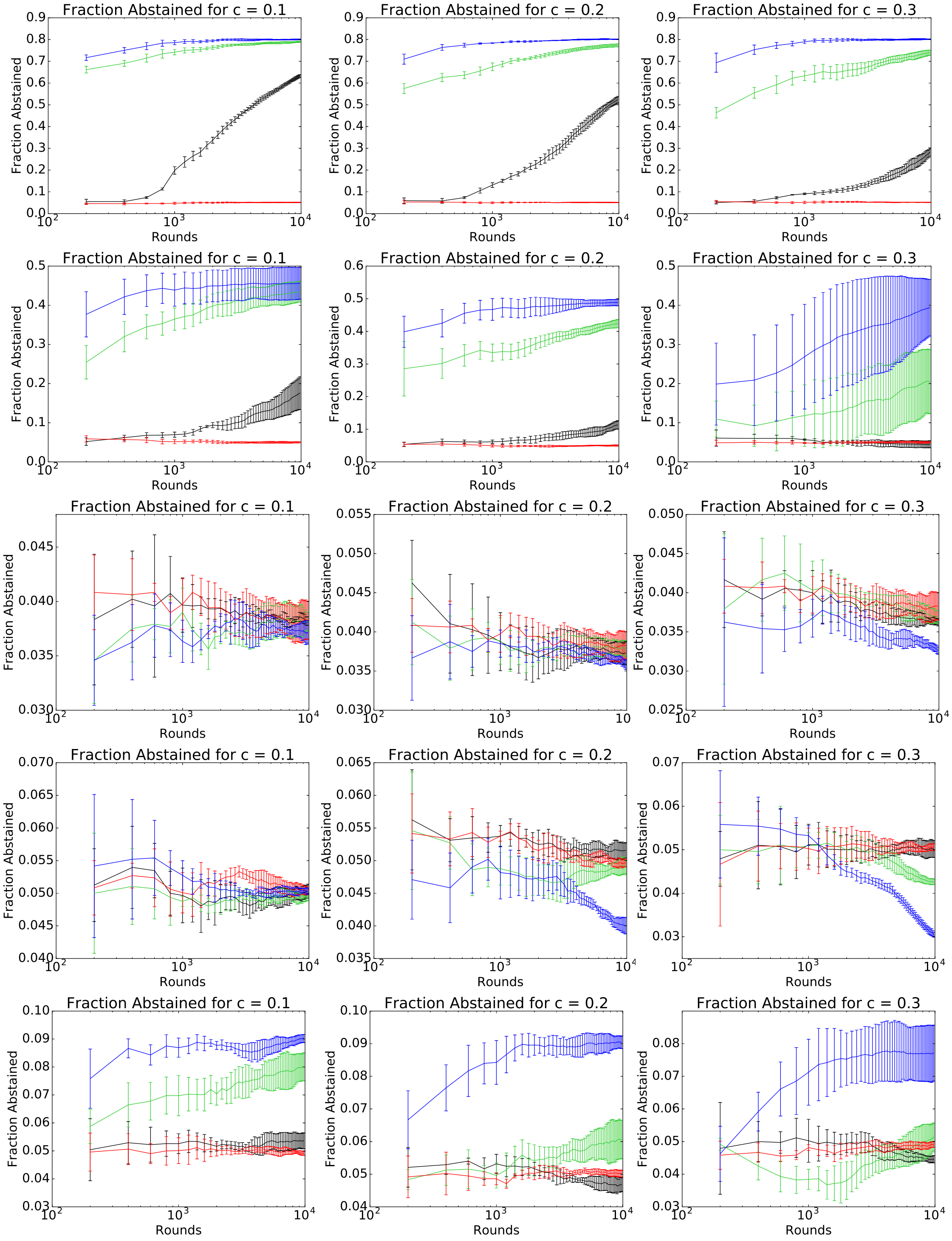

Figure 10: A graph of the averaged fraction of abstained points with standard deviations as a function of t (log scale) for UCB-GT, UCB-NT, UCB, and FS for different values of abstention costs. Each row is a dataset, starting from the top row we have: eye, cod-ran, synthetic, skin, and guide.

D.3. Average regret and fraction of abstention points for extreme abstention costs

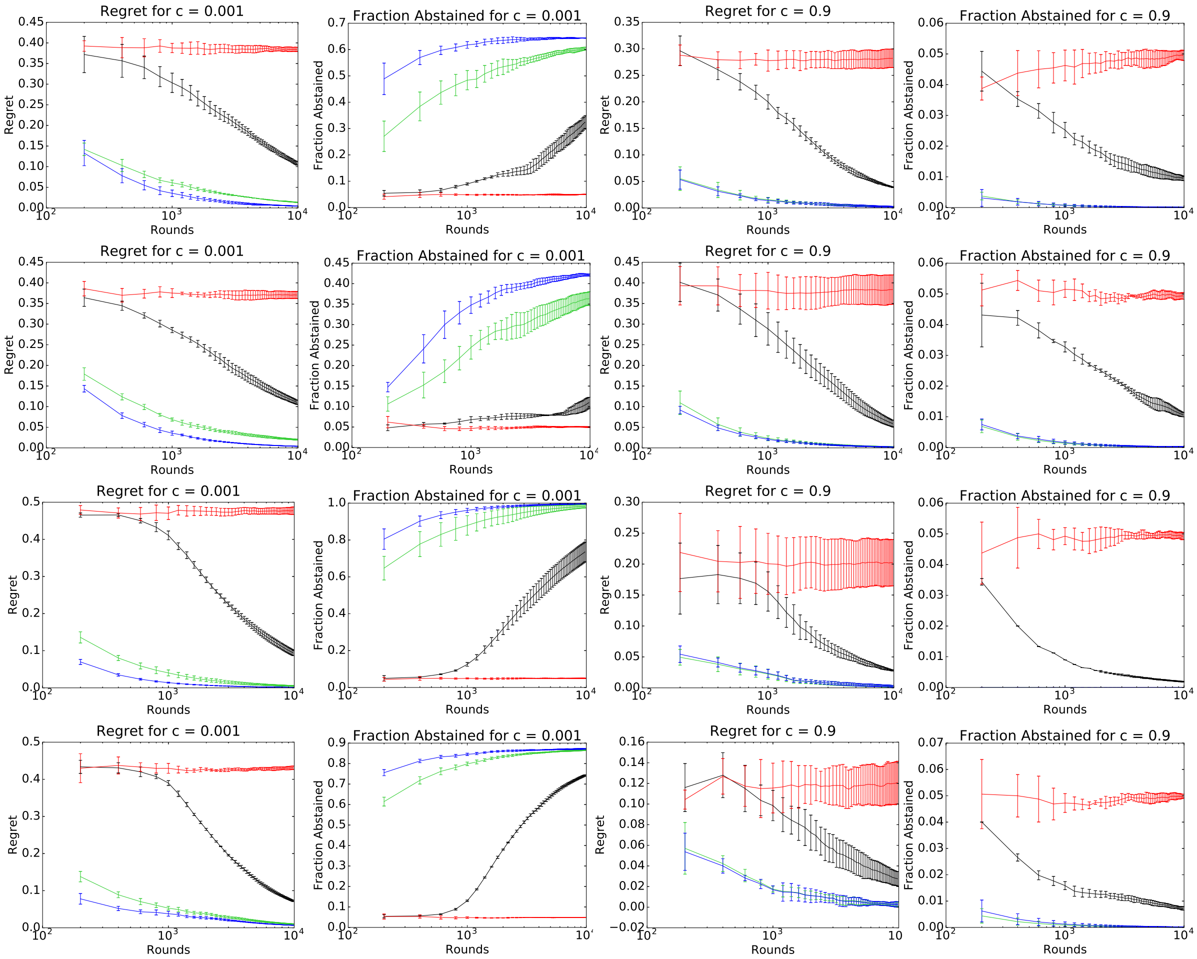

Figure 11: A graph of the averaged regret $R_t(\cdot)/t$ and fraction of points rejected with standard deviations as a function of t (log scale) for UCB-GT, UCB-NT, UCB, and FS for different values of abstention costs. The fraction of points decreases as the cost of abstention increases. The UCB-GT outperforms UCB-NT and UCB while approaching the performance of FS even at these extreme values of c. Each row is a dataset, starting from the top row we have: CIFAR, ijcnn, phishing, and covtype.

D.4. Average regret for confidence-based experts

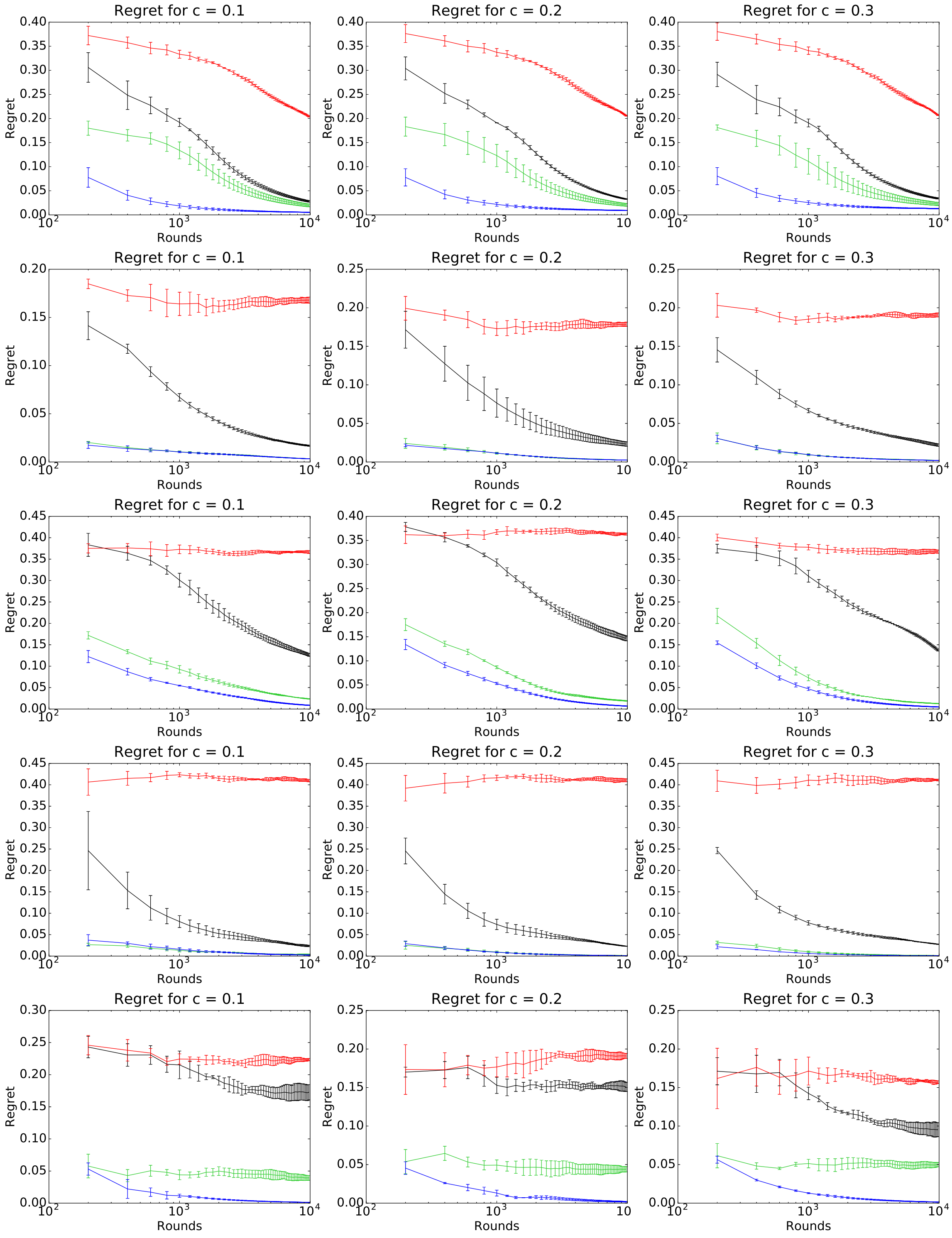

Figure 12: A graph of the averaged regret $R_t(\cdot)/t$ with standard deviations as a function of t (log scale) when using the confidence based experts for UCB-GT, UCB-NT, UCB, and FS. Each row is a dataset, starting from the top row we have: synthetic, skin, guide, ijcnn and CIFAR.

D.5. Average regret for a smaller set of experts

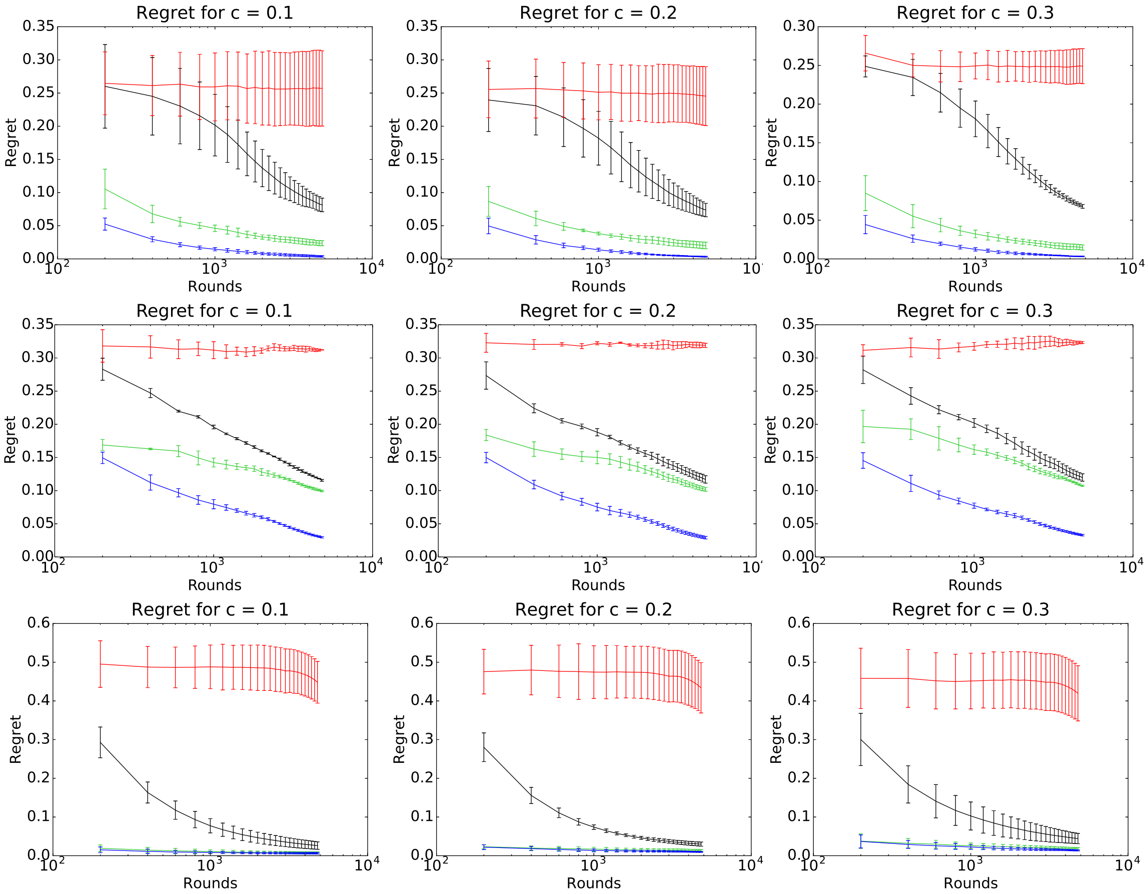

Figure 13: A graph of the averaged regret $R_t(\cdot)/t$ of abstained points with standard deviations as a function of t (log scale) for UCB-GT, UCB-NT, UCB, and FS for different values of abstention costs. Each row is a dataset, starting from the top row we have: guide, synthetic, and skin. We used $K = 500$ experts and $T = 5{,}000$ rounds in order to see the effect when changing the number of experts used.