

---

# Online Learning with Abstention

---

Corinna Cortes<sup>1</sup> Giulia DeSalvo<sup>1</sup> Claudio Gentile<sup>1,2</sup> Mehryar Mohri<sup>3,1</sup> Scott Yang<sup>\* 4</sup>

## 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.

<sup>1</sup>Google Research, New York, NY. <sup>2</sup>INRIA Lille Nord Europe.

<sup>3</sup>Courant Institute of Mathematical Sciences, New York, NY. <sup>4</sup>D. E. Shaw & Co., New York, NY. Correspondence to: Giulia DeSalvo <giuliad@google.com>.

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 CONTEXTUAL EXP3 (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  $h: \mathcal{X} \rightarrow \mathbb{R}$ , 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  $1_{yh(x) \leq 0}$ , 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  $r: \mathcal{X} \rightarrow \mathbb{R}$ , with  $r(x) \leq 0$  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  $r_{I_t}(x_t) \leq 0$  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

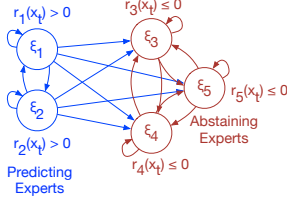


Figure 1: Feedback graph  $G_t^{\text{ABS}}$  for the scenario of online learning with abstention, with  $K = 5$ .

follows:

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

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  $\xi_i \rightarrow \xi_j$  is in  $E_t$  if the loss of expert  $\xi_j \in V$  is observed when expert  $\xi_i$  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  $\xi_j$ , since the label  $y_t$  is revealed to her. If instead she abstains at round  $t$  (i.e., if  $r_{I_t}(x_t) \leq 0$ ), then she only observes  $L(\xi_j, z_t)$  for those experts  $\xi_j$  that are abstaining in that round, that is, the set of  $j$  such that  $r_j(x_t) \leq 0$ , since for all such  $\xi_j$ , 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  $G_t^{\text{ABS}}$  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  $I_t$ s and, in the stochastic setting, the random draw of the  $z_t$ s.

In the stochastic setting, we will be mainly concerned with the case where  $\mathcal{E}$  is a finite set of experts  $\mathcal{E} = \{\xi_1, \dots, \xi_K\}$ . We then denote by  $\mu_j$  the expected loss of expert  $\xi_j \in \mathcal{E}$ ,  $\mu_j = \mathbb{E}_{z \sim \mathcal{D}}[L(\xi_j, z)]$ , by  $\mu^*$  the expected loss of the best expert,  $\mu^* = \min_{j \in [K]} \mu_j$ , and by  $\Delta_j$  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  $G_t^{\text{ABS}}$  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  $\eta$  over a set of  $K$  experts  $\xi_1, \dots, \xi_K$ . Then, the algorithm admits the following regret guarantee after  $T$  rounds:*

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

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

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

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

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

introduce some convex surrogate of the abstention loss (1), and work in the parametric space of  $(w, \theta)$  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  $f_\gamma$  is the 0/1-loss with slope  $1/(2\gamma)$  at the origin,  $f_\gamma(a) = \left(\frac{\gamma+a}{2\gamma}\right) 1_{|a| \leq \gamma} + 1_{a \geq 0} 1_{|a| > \gamma}$  (see Figure 2 (a)), and the class of experts  $\mathcal{E} = \{\xi = (h, r) \mid h, r: \mathcal{X} \subseteq \mathbb{R}^d \rightarrow [-1, 1]\}$ . 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  $\tilde{L}$  that upper bounds the abstention loss while maintaining the same feedback assumptions, namely the feedback graph given in Figure 1. One Lipschitz function  $\tilde{L}$  that precisely solves this problem is the following:

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

for  $\gamma \in (0, 1)$ .  $\tilde{L}(\xi, z)$  is plotted in Figure 2(b). Notice that this function is consistent with the feedback requirements of Section 2:  $r_{I_t}(x_t) \leq 0$  implies that  $\tilde{L}((h(x_t), r(x_t)), z_t)$  is known to the algorithm (i.e., is independent of  $y_t$ ) for all  $(h, r) \in \mathcal{E}$  such that  $r(x_t) \leq 0$ , while  $r_{I_t}(x_t) > 0$  gives complete knowledge of  $\tilde{L}((h(x_t), r(x_t)), z_t)$  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  $\tilde{L}(\xi, z_t), t \in [T]$ . The algorithm adaptively covers  $\mathcal{X}$  with balls of a fixed radius  $\varepsilon$ , 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*

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

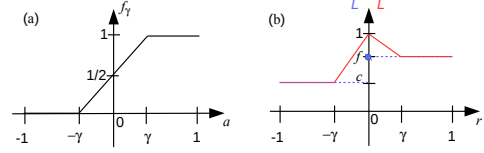


Figure 2: (a) The 0/1-loss function with slope  $1/(2\gamma)$  at the origin. (b) For a given value of  $x$  and margin  $a = -yh(x)$  (which in turn sets the value of  $f = f_\gamma(a) \in [0, 1]$ ), plots of the abstention loss  $L(a, r)$  (dotted blue curve), and the surrogate abstention loss  $\tilde{L}(a,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

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:

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

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

In the above,  $\tilde{\mathcal{O}}$  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  $G_t^{\text{ABS}}$ . 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  $\xi_j$  in the directed graph at time  $t$ , i.e., the set of vertices  $\xi_k$  destinations of an edge from  $\xi_j$ . Then, as with UCB-N, the algorithm updates, at



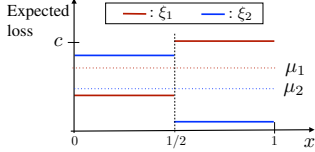


Figure 3: Illustration of the bias problem.

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  $G_t^{\text{ABS}}$  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  $G_t^{\text{ABS}}$ . 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  $G_t^{\text{ABS}}$  graphs. Nevertheless, we will show in Section 4.2 that UCB-NT does benefit from favorable guarantees, provided the feedback graph  $G_t^{\text{ABS}}$  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:  $\xi_1$  (red) and  $\xi_2$  (blue) with  $\mu_2 < \mu_1$  and  $\mathcal{X} = [0, 1]$  (see Figure 3). For  $x > \frac{1}{2}$ , the red expert  $\xi_1$  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  $\xi_2$ . Furthermore, it is abstaining and thus providing no updates for expert  $\xi_2$  (which is instead predicting). Hence, the algorithm continues to maintain an estimate of  $\xi_2$ '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  $x$ s and incurs a high regret.<sup>1</sup>

This simple example shows that, unlike the adversarial scenario (Section 3),  $G_t^{\text{ABS}}$ , 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

<sup>1</sup> 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.

---

#### ALGORITHM 1: UCB-NT

---

```

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

---

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  $\mu_j$  of the experts, which will allow us to apply standard concentration bounds in the proof of this algorithm.

#### 4.2. Time-varying graphs for UCB-NT

We now show that UCB-NT benefits from favorable guarantees, so long as the feedback graph  $G_t^{\text{ABS}}$  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  $s = 1, \dots, t-1$ , and  $i \in [K]$ , and let us denote this feedback graph by  $G_t$ , so as not to get confused with  $G_t^{\text{ABS}}$ . Under this assumption, we can derive strong regret guarantees for UCB-NT with time-varying graphs. Our guarantees are expressed in terms of the best sequence of *admissible  $p$ -partitionings* of graphs  $G_t$ . For  $p \in [K]$ , we say that  $(\mathcal{C}_{t,k})_{k \in [p]}$  is an admissible  $p$ -partitioning of  $G_t = (V, E_t)$  if  $V = \{\xi_j : j \in [K]\}$  is the union of  $p$  (disjoint) components  $\mathcal{C}_{t,k}$ , that is  $V = \bigcup_{k \in [p]} \mathcal{C}_{t,k}$ , and all vertices within a component  $\mathcal{C}_{t,k}$  are neighbors in  $G_t$  with reciprocal edges: if  $\xi_i$  and  $\xi_j$  are in  $\mathcal{C}_{t,k}$ , then we have  $\xi_i \in N_t(\xi_j)$  and  $\xi_j \in N_t(\xi_i)$ . Note that admissible  $p$ -partitionings are typically not unique since two neighbor vertices  $\xi_i$  and  $\xi_j$  may be placed in the same component or not. We denote by  $\mathcal{S}_p$  the set of all sequences  $((\mathcal{C}_{1,k})_{k \in [p]}, \dots, (\mathcal{C}_{T,k})_{k \in [p]})$  of admissible  $p$ -partitionings  $(\mathcal{C}_{t,k})_{k \in [p]}$  of graphs  $G_t$ . Moreover, for any sequence  $((\mathcal{C}_{1,k})_{k \in [p]}, \dots, (\mathcal{C}_{T,k})_{k \in [p]})$  in  $\mathcal{S}_p$ , we denote by  $\mathcal{C}_k$  the union of the components indexed by  $k$  over all

rounds:  $\mathcal{C}_k = \bigcup_{t \in [T]} \mathcal{C}_{t,k}$ .

**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:*

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

For a sequence  $\mathcal{S}_p$  made up of the same partition  $(\mathcal{C}_{1,k})_{k \in [p]}$  repeated  $T$  times, the theorem gives a bound on the regret based on this fixed partition, as it is the sum of  $p$  components, one per cluster  $\mathcal{C}_{1,k}$  in the partition. The minimum over  $(p, \mathcal{S}_p)$  then simply chooses the number of clusters  $p$  and the partitioning of  $V$  into  $p$  clusters having 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 also uses the newly defined notion of admissible  $p$ -partitionings 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  $p$ -partitioning for some small  $p \ll K$  (e.g. if the feedback graphs can be decomposed into a small number of components that are approximately fixed across time) for which  $\max_{j \in \mathcal{C}_k} \Delta_j \approx \min_{j \in \mathcal{C}_k} \Delta_j$ , then this bound can be up to a factor  $\frac{p}{K}$  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)  $K$ -partitioning that splits  $V$  into  $K$  singletons for all  $t$  is an admissible  $K$ -partitioning for all  $G_t$ 's. Furthermore, note that by construction, all vertices within the same component of an admissible  $p$ -partitioning are connected to one another. Thus, if the feedback graph is fixed throughout all rounds and we interpret the doubly-directed edges as edges of an undirected graph  $G_U$ , 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 clique 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  $G_t^{\text{ABS}}$  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*,  $G^{\text{SUB}}$ , 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  $\xi_i$  is a subset of that of  $\xi_j$ , then, if  $\xi_i$  is selected at some round  $t$  and is abstaining, so is  $\xi_j$ . For an example, see  $\xi_i$  and  $\xi_j$  in Figure 4 (top). Crucially, this implication holds regardless of the particular input point  $x_t$  received in the region of abstention of  $\xi_i$ . Thus, the set of vertices of  $G^{\text{SUB}}$  is  $\mathcal{E}$ , and  $G^{\text{SUB}}$  admits an edge from  $\xi_i$  to  $\xi_j$ , iff  $\{x \in \mathcal{X}: r_i(x) \leq 0\} \subseteq \{x \in \mathcal{X}: r_j(x) \leq 0\}$ . Since  $G^{\text{SUB}}$  does not vary with time, it trivially verifies the condition of the previous section. Thus, UCB-NT run with  $G^{\text{SUB}}$  admits the regret guarantees of Theorem 3, where the admissible  $p$ -partitionings are those of fixed graph  $G^{\text{SUB}}$ .

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

**Proposition 1** *Assume that UCB-NT is run with feedback graphs  $G_t$  that are not subsets of  $G^{\text{SUB}}$ . Then, there exists a family of predictors  $\mathcal{H}$ , a Lipschitz loss function  $\ell$  in (1), and a distribution  $\mathcal{D}$  over  $z_t$ 's 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  $G^{\text{SUB}}$ . 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  $G^{\text{SUB}}$ ? 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  $G^{\text{SUB}}$  as a subgraph. We will show how certain types of edges can be safely added to  $G^{\text{SUB}}$  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

**ALGORITHM 2: UCB-GT**


---

```

for  $t \geq 1$  do
    RECEIVE( $x_t$ );
     $\xi_{I_t} \leftarrow \operatorname{argmin}_{\xi_i \in \mathcal{E}} \{\widehat{\mu}_{i,t-1} - S_{i,t-1}\}$ ,
    where  $S_{i,t-1}$  is as in Algorithm 1;
    for  $\xi_i \in \mathcal{E}$  do
        if  $\widehat{p}_{i,t-1}^{t-1} \leq \gamma_{i,t-1}$  then  $Q_{i,t} \leftarrow Q_{i,t-1} + 1$ ;
        if  $r_{I_t}(x_t) \leq 0 \wedge r_i(x_t) > 0$  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
    
```

---

$G^{\text{SUB}}$  with edges from  $\xi_j$  to  $\xi_i$  where the subset property  $\{x: r_j(x) \leq 0\} \subseteq \{x: r_i(x) \leq 0\}$  may not hold, but where the implication  $(r_j(x) \leq 0 \Rightarrow r_i(x) \leq 0)$  holds with high probability over the choice of  $x \in \mathcal{X}$ , that is, the region  $\{x: r_j(x) \leq 0 \wedge r_i(x) > 0\}$  admits low probability. Of course, adding such an edge  $\xi_j \rightarrow \xi_i$  can cause the estimation bias of Section 4.1. But, if we restrict ourselves to cases where  $p_{j,i} = \mathbb{P}[r_j(x) \leq 0 \wedge r_i(x) > 0]$  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  $G^{\text{SUB}}$ , 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  $\widehat{p}_{j,i}^{t-1} := \frac{1}{t-1} \sum_{s=1}^{t-1} 1_{r_j(x_s) \leq 0, r_i(x_s) > 0}$ . At time  $t$ , if expert  $\xi_j$  is selected, we update expert  $\xi_i$  if the condition  $\widehat{p}_{j,i}^{t-1} \leq \gamma_{i,t-1}$  holds with  $\gamma_{i,t-1} = \sqrt{5Q_i(t-1) \log(t) / ((K-1)(t-1))}$ . If the expert  $\xi_{I_t}$  chosen abstains while expert  $\xi_j$  predicts and satisfies  $\widehat{p}_{I_t,j}^{t-1} \leq \gamma_{j,t-1}$ , 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  $\xi_j$ :  $N_t(j) = \{\xi_i \in \mathcal{E}: \widehat{p}_{j,i}^{t-1} \leq \gamma_{i,t-1}\}$ . Let  $\mathcal{S}_p = (\mathcal{C}_{t,k})_{t \in [T], k \in [p]}$  denote any sequence of admissible  $p$ -partitionings of these feedback graphs, then 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  $N_t(j) = \{\xi_i \in \mathcal{E}: \widehat{p}_{j,i}^{t-1} \leq \gamma_{i,t-1}\}$ . Then, the regret of UCB-GT is bounded as follows:*

$$\mathcal{O}\left(\mathbb{E}\left[\min_{p, \mathcal{S}_p} \sum_{k \in [p]} \frac{\max_{j \in \mathcal{C}_k} \Delta_j}{\min_{j \in \mathcal{C}_k} \Delta_j} \log(T) + K\right]\right).$$

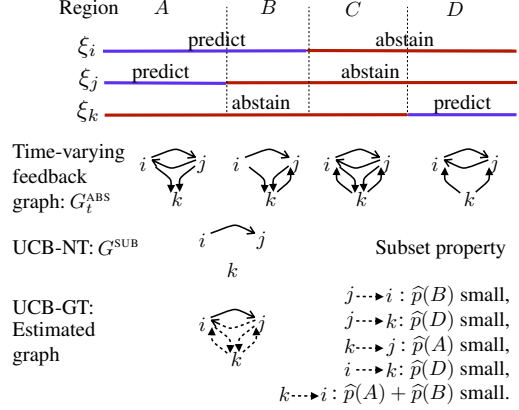


Figure 4: The top row shows three experts  $\xi_i, \xi_j$ , and  $\xi_k$  on a one-dimensional input space marked by their prediction and abstention regions. Below each region, the time-varying graph  $G_t^{\text{ABS}}$  is shown. To avoid the bias problem affecting the graphs  $G_t^{\text{ABS}}$ , one option is to use  $G^{\text{SUB}}$ . Yet, as illustrated,  $G^{\text{SUB}}$  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  $\xi_j$  to  $\xi_i$  if the probability of the region where  $\xi_j$  is abstaining but  $\xi_i$  is predicting is (estimated to be) small.

Since the graph  $G_t$  of UCB-GT has more edges than  $G^{\text{SUB}}$ , it admits more admissible partitionings than  $G^{\text{SUB}}$ , which leads to a more favorable guarantee than that of UCB-NT run with  $G^{\text{SUB}}$ . 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  $G^{\text{SUB}}$ , 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  $\xi_j$  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, ijcn, 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

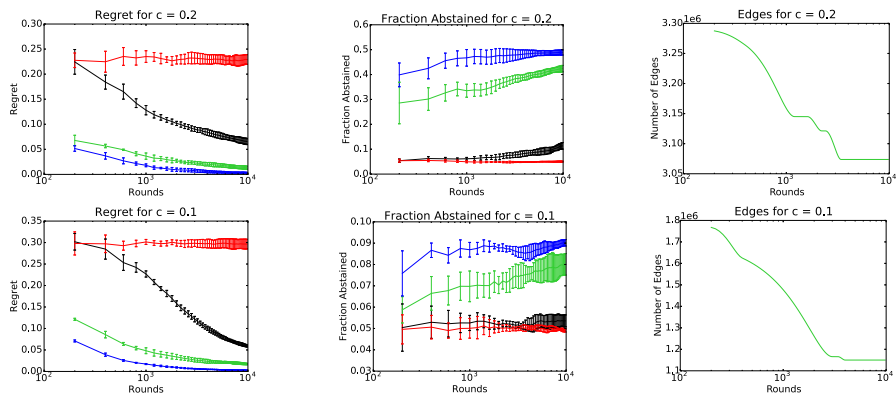


Figure 5: From the left, we show graphs of the average regret  $R_t(\cdot)/t$ , fraction of points the chosen expert abstained on, and the number of edges of the feedback graph as a function of  $t$  (log-scale) for UCB-GT, UCB-NT, UCB, and FS. Top row is the results for `cod-rna` for cost  $c = 0.2$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

projections as features, and a synthetic dataset of points drawn according to the uniform distribution in  $[-1, 1]^2$ . For 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  $(0, \frac{\sqrt{d}}{20}, \frac{2\sqrt{d}}{20}, \dots, \sqrt{d})$ . 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.



## 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.
- 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.
- Sayed, 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.