Delaytron: Efficient Learning of Multiclass Classifiers with Delayed Bandit Feedbacks

Naresh Manwani  
Machine Learning Lab  
IIIT Hyderabad  
India  
narehs.manwani@iiit.ac.in

Mudit Agarwal  
Machine Learning Lab  
IIIT Hyderabad  
India  
arawalmudit9876@gmail.com

Abstract—This paper presents an online algorithm Delaytron for learning multiclass classifiers using delayed bandit feedback. At the t-th round, the algorithm observes an example x_t and predicts a label ŷ_t and receives the bandit feedback I[ŷ_t = y_t] only d_t rounds later. When t + d_t > T, we consider the feedback for the t-th round missing. The sequence of feedback delays \{d_t\}_{t=1}^T is unknown to the algorithm. For the K-class classification problem, we show that the proposed algorithm achieves regret of \(O(\sqrt{\frac{2K}{T} T + \left(2 + \frac{T^2}{R^2 L^2} \right) \sum_{t=1}^T d_t})\) when the loss for each missing sample is upper bounded by L. In the case when the loss for missing samples is not upper-bounded, the regret achieved by Delaytron is \(O(\sqrt{\frac{2K}{T} T + 2 \sum_{t=1}^T d_t + |M| T})\) where M is the set of missing samples in T rounds. These bounds were achieved with a constant step size which requires the knowledge of T and \(\sum_{t=1}^T d_t\). When T and \(\sum_{t=1}^T d_t\) are unknown, we use a doubling trick for online learning and propose Adaptive Delaytron. We show that Adaptive Delaytron achieves a regret bound of \(O(\sqrt{T + \sum_{t=1}^T d_t})\). We experimentally show that the proposed approach can learn efficient classifiers even with delayed bandit feedback, and the accuracy does not degrade much due to delays in feedback.

Index Terms—Online learning, classification, bandit feedback, feedback delay, regret analysis.

I. INTRODUCTION

Learning classifiers using bandit feedback is a well-addressed problem in the machine-learning community. The learner can only get feedback on whether the predicted label is correct in the bandit feedback setting. In [12], the Banditron algorithm is proposed to learn with bandit feedback. When the data is linearly separable, Banditron achieves a mistake bound of \(O(\sqrt{T})\) in the expected sense. In the general case, Banditron makes \(O(T^{2/3})\) mistakes. On the other hand, Newton [9], which uses the online Newton method on a strongly convex objective function (adding regularization term with the loss function), achieves \(O(\log T)\) regret bound in the best case and \(O(T^{2/3})\) in the worst case. An exact passive-aggressive approach is proposed to learn classifiers in the bandit setting in [2]. A second-order algorithm is discussed in [4] which achieves \(O(\frac{T}{\log T})\) regret. Banditron algorithm has been extended in the noisy bandit feedback setting in [1] and shown to achieve \(O(T^{2/3})\) mistake bound in the worst case. A dilute bandit feedback setting is proposed in [3] in which the algorithm predicts a subset of labels and gets the feedback on whether the actual label lies in that set. It is shown that even with this weaker bandit feedback, the algorithm can learn good classifiers [3].

The above approaches assume that the feedback for trial t is received in the same round—however, it is not the case in many practical situations. Consider the example of sponsored advertising in which the user’s queries are answered as a list of product ads to maximize the clicks (by considering the query and the user’s history and buying patterns) [10]. Users’ clicks on a particular ad might indicate the user’s interest in that product. However, the user may take their own sweet time to click on an ad, and sometimes it may not. This causes a delay in the feedback on whether the user liked some ad. Meanwhile, the system still serves ads to other users and does not wait to know previous users’ click outcomes. This situation also arises in the healthcare system where the feedback of a specific treatment on a patient can be received only after some time or delay [16]. This delay is natural as it takes time to observe the effects of the treatment. However, the healthcare professional still needs to treat other patients for the same problem before watching the treatment outcome of the previous patients. In finance, online learning algorithms used to manage portfolios always suffer from information and transaction delays from the market [14]. Financial firms invest a massive amount of effort in minimizing these delays.

Adversarial multi-arm bandits (with k-arms) were first studied in [6] under the assumption of fixed delay d. In T-trials, they gave a lower bound of \(\Omega(\max(\sqrt{kT}, \sqrt{T d T \log k}))\) for \(d \leq \frac{T}{\log k}\) and a matching upper bound of \(O(\sqrt{dT \log k} + \sqrt{kT \log k})\). Bistritz et al. [5] proposed a variant of exp3 algorithm that can handle non-uniform delays and achieves \(O(\sqrt{kT \log(k)} + \sqrt{\sum_{t=1}^T d_t \log(k)})\) regret bound under the assumption that T and \(\sum_{t=1}^T d_t\) are known in advance. Bistritz et al. [5] provide a doubling scheme that achieves an \(O(\sqrt{k^2 T \log(k)} + \sqrt{\sum_{t=1}^T d_t \log(k)})\) regret bound when T and \(\sum_{t=1}^T d_t\) are unknown. Stochastic gradient descent on convex objective function with delayed feedback is explored in [17]. In [17], it is assumed that the convex function corresponding to time t is revealed after a delay of d_t only and proposed an approach...
which achieves a regret bound of $O(\sqrt{\sum_{t=1}^{T} d_t})$.

No work deals with multiclass classification with delayed bandit feedback. This paper proposes an approach that efficiently learns classifiers using delayed bandit feedback. Following are the key features of our proposed method.

1) The proposed approach achieves a regret bound of $O\left(\sqrt{\sum_{t=1}^{T} d_t} + \left[\frac{2K^2}{\gamma} \frac{1}{T^2} + \left(2 + \frac{2R^2}{\gamma \|W\|_F^2}\right) \sum_{t=1}^{T} d_t\right]\right)$ for constant step size. Here, $K$ is the number of classes, $\mathcal{M}$ is the set of samples for which the feedback is not received till $T$ and $L$ is the upper bound on the loss for samples in $\mathcal{M}$.

2) When there is no boundedness assumption on the loss, using 0-1 loss ($\epsilon$), we show experimentally that the proposed algorithms (Delaytron and Adaptive Delaytron) can learn efficient classifiers using delayed bandit feedback.

II. PROPOSED APPROACH: DELAYTRON

Here, the goal is to learn a function $g : \mathcal{X} \rightarrow [K]$ which given an example $x \in \mathcal{X} \subseteq \mathbb{R}^d$ assigns a label in the set $[K] = \{1, \ldots, K\}$. A linear multi-class classifier is modeled using a weight matrix $W \in \mathbb{R}^{K \times d}$ as $g(x) = \arg\max_{j \in [K]} \langle W, x \rangle$. To learn the weight matrix $W$, we use the training set $\{(x_1, y_1), \ldots, (x_T, y_T)\}$ where $(x_i, y_i) \in \mathcal{X} \times [K]$, $\forall i \in [T]$. Performance of classifier $g$ is measured using 0-1 loss ($l = 1$) loss, where $\sum_{i=1}^{T} l_0(W, (x_i, y_i)) = \|g(x_i) \neq y_i\|$. In practice, convex surrogates of $l_0$ are used to learn the classifier. One such surrogate is multiclass hinge loss [7] described as

$$l_H(W, (x_i, y_i)) = \max(0, 1 - \langle W, x_i \rangle, y_i) \max_{j \neq y_i} \langle W, x_i \rangle).$$

At round $t$, the algorithm receives an example $x_t$. It then finds $\hat{y}_t$ as $\hat{y}_t = \arg\max_{j \in [K]} \langle W, x_t \rangle$. Then, it defines a distribution over the classes as $P_t(r) = (1 - \gamma)\|r = \hat{y}_t\| + \frac{\gamma}{K}$, $r \in [K]$. It samples $\hat{y}_t$ using distribution $P_t$ and predicts $\hat{y}_t$. The algorithm gets to know the bandit feedback ($\|\hat{y}_t = y_t\|$) of predicting $\hat{y}_t$ for round $t$ at the end of the $t + d_t - 1$ round. In other words, the algorithms observes a delay of $d_t \geq 1$ to observe the feedback $\|\hat{y}_t = y_t\|$. So, the feedback is available at the beginning of round $t + d_t$. Note that $y_t$ is the true label of the example $x_t$. The goal of the algorithm is to minimize the number of mistakes $M = \sum_{t=1}^{T} l_0(W, x_t, y_t)$.

Let $\mathcal{S}$ denote the set of feedbacks received at the beginning of round $t$. So, $r \in \mathcal{S}$ means that the feedback of predicting $\hat{y}_t$ at round $r < t$ is received at round $t$. Since the algorithm runs for $T$ rounds, all feedbacks for which $t + d_t > T$ are never received. These feedbacks are called missing samples, and $\mathcal{M}$ denotes the set of missing samples. At round $t$, the algorithm considers all the feedbacks received at time $t$ and updates the parameters as $W^{t+1} = W^t + \eta \sum_{s \in \mathcal{S}} U_s$, where $\eta > 0$ is the step size. Also, for all $r \in [K], j \in [d], \hat{U}_{r,j}$ is defined as

$$\hat{U}_{r,j} = x_t(j) \left[\frac{\|\hat{y}_t = y_t\|}{P_t(r)} - \frac{\|\hat{y}_t = r\|}{P_t(r)}\right].$$

Note that $\mathbb{E}_{y_t \sim P_t}[\hat{U}_{r,j}] = U_{r,j}$ [12], where matrix $U^t$ is negative of the gradient of the loss $\{1 - \langle W, x_t \rangle, y_t\}$ at $\hat{U}_{r,j}$ at $\hat{W}^t$. The approach is described in detail in Algorithm 1.

Algorithm 1 Delaytron

Input: $\gamma \in (0, 0.5)$, Step size $\eta > 0$

Initialize: Set $W^1 = 0 \in \mathbb{R}^{K \times d}$

1: for $t = 1, \ldots, T$ do
2: Receive $x_t \in \mathbb{R}^d$.
3: Set $\hat{y}_t = \arg\max_{j \in [K]} \langle W, x_t \rangle$.
4: Set $P_t(r) = (1 - \gamma)\|r = \hat{y}_t\| + \frac{\gamma}{K}$, $r \in [K]$.
5: Randomly sample $\hat{y}_t$ according to $P_t$ and predict $\hat{y}_t$.
6: Obtain a set of delayed feedbacks $\|\hat{y}_t = y_t\|$ for all $s \in \mathcal{S}$, where $y_t$ is the prediction at round $s$ and $y_s$ is the corresponding ground truth.
7: for $s \in \mathcal{S}$ do
8: For all $r \in [K]$ and $j \in [d]$, compute $\hat{U}_{s,j}$ as below:
9: $\hat{U}_{s,j} = x_t(j) \left[\frac{\|\hat{y}_t = y_t\|}{P_t(r)} - \frac{\|\hat{y}_t = r\|}{P_t(r)}\right]$.
10: end for
11: Update: $W^{t+1} = W^t + \eta \sum_{s \in \mathcal{S}} \hat{U}^s$.

III. ANALYSIS

We split the sum of gradients in a single round and applied them one by one. For each $s \in \mathcal{S}$, let $\mathcal{S}_{t,s} = \{q \in \mathcal{S} : q < s\}$, which is the set of bandit feedback samples that the algorithm uses before the bandit feedback from round $s$ is used. Let $F_t = \sigma(\{y_t \mid s + d_t \leq t\})$, which is generated from all the actions for which the feedback was received up to round $t$. The performance of the algorithm is measured using the regret as follows.

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} l_H(W_t, (x_t, y_t)) - \sum_{t=1}^{T} l_H(W^*, (x_t, y_t))\right]$$

where $W^* = \arg\min_{W} \sum_{t=1}^{T} l_H(W, (x_t, y_t))$. We first show the regret bound for the examples other than missing samples. We use the following lemma to analyze the "delay term" contribution to the expected regret.

Lemma 1. [5] Let $d_t$ be the delay in the bandit feedback corresponding to round $t$. Let $\mathcal{S}_t$ be the set of feedbacks received at time $t$ and define $\mathcal{S}_{t,s} = \{q \in \mathcal{S}_t : q < s\}$ which is the set of feedback samples used before the feedback from round $s$ is used. Let $\mathcal{M}$ be the set of missing samples, then

$$\sum_{t=1}^{T} \sum_{s \in \mathcal{S}_t} |\mathcal{S}_{t,s}| + \sum_{r=1}^{T} |\mathcal{S}_t| \leq 2 \sum_{t \notin \mathcal{M}} d_t.$$
Theorem 1. Let \( \{x_i\}_{i=1}^T \) be the sequence of examples observed by Algorithm 1. Let \( \|y_i - y_t\| \) be the bandit feedback corresponding to round \( t \), which is observed only after a delay of \( d_t \). Let \( \|x_t\| \leq R, \forall t \in [T] \). Then for any \( W \in \mathbb{R}^{K \times d} \), the regret achieved by Delaytron on examples for which the feedback is received before \( T \) is as follows:

\[
\mathbb{E} \left[ \sum_{t \in \mathcal{M}} l_H(W^t, (x_t, y_t)) - \sum_{t \in \mathcal{M}} l_H(W, (x_t, y_t)) \right] \\
\leq \frac{1}{2\eta} \|W\|_F^2 + \frac{\eta K R^2}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t \right]
\]

We analyze these results further for three different cases as follows.

A. Case 1: No Delay

When there is no delay, it becomes the standard learning setting with bandit feedback as discussed in [12]. Here, we set \( \mathcal{M} \) is empty and \( d_t = 1, \forall t \in [T] \). Thus, \( \sum_{t \in \mathcal{M}} d_t = \sum_{t=1}^T d_t = T \). Using \( \eta = \frac{|\mathcal{M}|}{\sqrt{2KdR}} \), the regret bound becomes

\[
\mathbb{E} \left[ \sum_{t=1}^T l_H(W^t, (x_t, y_t)) - \sum_{t=1}^T l_H(W, (x_t, y_t)) \right] \\
\leq 6R\|W\|_F \sqrt{\frac{KT}{\gamma}}.
\]

The regret achieved in this case is \( O(\sqrt{T}) \). Note that when the data is linearly separable, then the mistake bound is also \( O(\sqrt{T}) \). A similar mistake bound has been achieved in [12] for the linearly separable case.

B. Case 2: Bounded Loss for Missing Samples

Here, we assume that the loss \( l_H(W, (x_t, y_t)) \) is upper bounded for all the missing samples. Which means, \( l_H(W, (x_t, y_t)) \leq L, \forall t \in \mathcal{M} \). This seems a fair assumption, as the bandit feedback for such examples is never revealed.

Theorem 2. Let \( \{x_i\}_{i=1}^T \) be the sequence of examples observed by Algorithm 1. Let \( \|y_i - y_t\| \) be the bandit feedback corresponding to round \( t \), which is observed only after a delay of \( d_t \). Let \( \|x_t\| \leq R, \forall t \in [T] \) and \( \mathcal{M} = \{t \in [T] \mid t + d_t > T\} \). Let \( l_H(W, (x_t, y_t)) \leq L, \forall t \in \mathcal{M}, \forall W \in \mathbb{R}^{K \times d} \). Then for any \( W \in \mathbb{R}^{K \times d} \), the regret achieved by Delaytron is as follows:

\[
\mathcal{R}(T) \leq R\|W\|_F \left[ \frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t \right] + L|\mathcal{M}| \right]
\]

Our bound does not account for the delays of the missing samples, and it only depends on \( \sum_{t \in \mathcal{M}} d_t \). Moreover, counting delays that go beyond \( T \) is redundant. It is worth noting that \( m = |\mathcal{M}| \) contribute at least \( m(m + 1)/2 \) to the sum of delays \( \sum_{t=1}^T d_t \). It happens in the following scenario. Feedback of round \( T \) is delayed by one, round \( (T-1) \) feedback is delayed by two, and so on \( m \) times. Using this idea, we get regret bound independent of \( \mathcal{M} \).

Corollary 1. Let \( \{x_i\}_{i=1}^T \) be the sequence of examples observed by Algorithm 1. Let \( \|y_i - y_t\| \) be the bandit feedback corresponding to round \( t \), which is observed only after a delay of \( d_t \). Let \( \|x_t\| \leq R, \forall t \in [T] \) and \( \mathcal{M} = \{t \in [T] \mid t + d_t > T\} \). Let \( l_H(W, (x_t, y_t)) \leq L, \forall t \in \mathcal{M}, \forall W \in \mathbb{R}^{K \times d} \). Then for any \( W \in \mathbb{R}^{K \times d} \), the regret achieved by Delaytron is upper bounded as follows:

\[
\mathcal{R}(T) \leq R\|W\|_F \left[ \frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t=1}^T d_t \right] + L|\mathcal{M}| \right]
\]

The regret bound above is independent of \( \mathcal{M} \). However, it depends on \( T \) and \( \sum_{t=1}^T d_t \).

C. Without Boundedness Condition on the Loss for Missing Samples

Here, we do not assume that the loss for the missing samples is upper bounded by a fixed quantity. Now, let us find regret on missing samples.

Theorem 3. Let \( \{x_i\}_{i=1}^T \) be the sequence of examples observed by Algorithm 1. Let \( \|y_i - y_t\| \) be the bandit feedback corresponding to round \( t \), which is observed only after a delay of \( d_t \). Let \( \|x_t\| \leq R, \forall t \in [T] \) and \( \mathcal{M} = \{t \in [T] \mid t + d_t > T\} \). Then for any \( W \in \mathbb{R}^{K \times d} \), the regret achieved by Delaytron is as follows.

\[
\mathcal{R}(T) \leq R\|W\|_F \left[ \frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}| T \right] \right]
\]

We get \( O \left( \sqrt{\frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}| T} \right) \) regret bound in the case when we don’t have any boundedness assumption on the loss incurred on missing samples.

Corollary 2. Let \( \{x_i\}_{i=1}^T \) be the sequence of examples observed by Algorithm 1. Let \( \|y_i - y_t\| \) be the bandit feedback corresponding to round \( t \), which is observed only after a delay of \( d_t \). Let \( \|x_t\| \leq R, \forall t \in [T] \) and \( \mathcal{M} = \{t \in [T] \mid t + d_t > T\} \). Then for any \( W \in \mathbb{R}^{K \times d} \), the regret achieved by Delaytron is as follows:

\[
\mathcal{R}(T) \leq O \left( R\|W\|_F \left[ \frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}| T \right] \right] \right)
\]

IV. ADAPTIVE DELAYTRON FOR UNKNOWN \( T \) AND \( \sum_{t=1}^T d_t \)

The fixed step size \( \eta \) used in Delaytron requires the knowledge of \( T \) and \( \sum_{t \in \mathcal{M}} d_t \). Using the doubling trick, we can make it independent of \( T \) and \( \sum_{t \in \mathcal{M}} d_t \). However, the usual doubling trick [15] does not work with delays. We use the doubling trick proposed in [5], where \( T \) and \( \sum_{t \in \mathcal{M}} d_t \) are unknown. We define \( m_t \) as the number of samples for which bandit feedback is missing till round \( t \). The idea is to start a new epoch every time \( \sum_{s=1}^{m_t} m_s \) (that tracks \( \sum_{s=1}^{m_t} d_s \) doubles. We define the \( e \)-th epoch as \( T_e = \{t \mid 2^{e-1} \leq \sum_{s=1}^t m_s < 2^e \} \).
$\mathcal{T}_e$ is the set of consecutive rounds in which the sum of delays is within a given interval. During the $e$-th epoch, the algorithm uses the step size $\eta_e = \sqrt{\frac{1}{2T_e}}$. The resulting approach is called Adaptive Delaytron and is discussed in detail in Algorithm 2.

**Theorem 4.** Let $\{x_t\}_{t=1}^T$ be the sequence of examples observed by Algorithm 1. Let $[\hat{y}_t = y_t]$ be the bandit feedback corresponding to round $t$, which is observed only after a delay of $d_t$. Let $\|x_t\|_2 \leq R$, $\forall t \in [T]$ and $\mathcal{M} = \{t \in [T] \mid t + d_t > T\}$. Let $l_H(W(x_t, y_t)) \leq L$, $\forall t \in \mathcal{M}, \forall W \in \mathbb{R}^{K \times d}$. Then for any $W \in \mathbb{R}^{K \times d}$, the regret achieved by Adaptive Delaytron is as follows:

$$\mathcal{R}(T) \leq 10 \left[ \frac{1}{2} \|W\|_F^2 + \frac{KR^2}{\gamma} + L \right] T + \sum_{t=1}^T d_t$$

**Proof.** Define $\mathcal{M}_e$ as the set of bandit feedbacks in epoch $e$ that are not received within epoch $e$. Denote by $T_e = \max \{T_e \mid t \in \mathcal{M}_e \}$ as the last round in $\mathcal{M}_e$. Note that $\mathcal{T}_e$ is also set of consecutive rounds from $T_{e-1} + 1$ to $T_e$. Every round $t \notin \mathcal{M}_e$ such that $t \in \mathcal{M}_e$ contributes exactly $d_t$ to $\sum_{s = T_{e-1} + 1}^{T_e} m_s$, since the $t$-th feedback is missing for $d_t$ rounds sometime between $T_{e-1} + 1$ and $T_e$. Therefore, $\sum_{t \notin \mathcal{M}_e, t \notin \mathcal{M}_e} d_t \leq \sum_{s = T_{e-1} + 1}^{T_e} m_s \leq 2^{-e-1}$.

Where the last inequality uses the fact that if $\sum_{s = T_{e-1} + 1}^{T_e} m_s > 2^{-e-1} \geq 2^{-e-1} + 2^{-e-1} = 2^e$, then epoch $e + 1$ should have been already started. We apply Theorem 1 separately on each epoch, which gives an upper bound on the regret of epoch $e$ as follows.

$$R_e := \mathbb{E} \left[ \sum_{t \in \mathcal{T}_e} l_H(W_t^e, (x_t, y_t)) - \sum_{t \in \mathcal{T}_e} l_H(W_t, (x_t, y_t)) \right] \leq \frac{1}{2\eta_e} \left\| W \right\|_F^2 + \frac{\eta_e KR^2}{\gamma} \left[ \frac{T_e}{2} + 2 \sum_{t \in \mathcal{T}_e, t \notin \mathcal{M}_e} d_t \right] + L |\mathcal{M}_e|$$

Now we find the largest set $\mathcal{M}_e$ such that $\sum_{s = T_{e-1} + 1}^{T_e} m_s \leq 2^{-e-1}$ is still possible. The cheapest way to that is when the feedback from round $T_e$ is delayed by $1$ (contributing $1$ to $\sum_{s = T_{e-1} + 1}^{T_e} m_s$), feedback from round $T_e - 1$ is delayed by $2$ (contributing $2$ to $\sum_{s = T_{e-1} + 1}^{T_e} m_s$) and so on. This gives $\sum_{s = T_{e-1} + 1}^{T_e} \frac{\left| \mathcal{M}_e \right|}{e} = \frac{\left| \mathcal{M}_e \right| \left( \left| \mathcal{M}_e \right| + 1 \right)}{2} \leq 2^{e-1}$. This can happen if $\left| \mathcal{M}_e \right| \leq 2\frac{e}{2}$. By choosing $\eta_e = 2^{-\frac{e}{2}}$, we obtain $R_e \leq \frac{1}{2\eta_e} \left\| W \right\|_F^2 + \frac{2\eta_e KR^2}{\gamma} \left[ \frac{T_e}{2} + 2 \right] + L 2^{\frac{e}{2}} \leq 2^{\frac{e}{2}} \left\| W \right\|_F^2 + \frac{KR^2}{\gamma} + L + 2KR^2 T_e 2^{-\frac{e}{2}}$. Let the last epoch be denoted as $E$. Then, $\mathbb{E}[\mathcal{R}(T)] = \sum_{e=1}^E R_e$ can be bounded as

$$\mathbb{E}[\mathcal{R}(T)] \leq \sum_{e=1}^E \left\{ 2^{\frac{e}{2}} \left[ \frac{1}{2} \left\| W \right\|_F^2 + \frac{KR^2}{\gamma} + L \right] + \frac{2KR^2T_e}{\gamma} 2^{-\frac{e}{2}} \right\}$$

$$= \frac{2^{E/2} - 1}{2 \sqrt{2} - 1} \left[ \frac{1}{2} \left\| W \right\|_F^2 + \frac{KR^2}{\gamma} + L \right] + \frac{2KR^2}{\gamma} \sum_{e=1}^E T_e 2^{-\frac{e}{2}}$$

The maximum of $\sum_{e=1}^E T_e 2^{-\frac{e}{2}}$ (subject to $\sum_{e=1}^E T_e = T$) is achieved when $E = \lceil \log_2 T \rceil$ and $e$-th epoch having length $2^e$. So, $\sum_{e=1}^E T_e 2^{-\frac{e}{2}} \leq \sum_{e=1}^{\lceil \log_2 T \rceil} 2^{\frac{e}{2}} \leq \sqrt{2^{\lceil \log_2 T \rceil} - 1} \leq 5\sqrt{T}$. We also note that $\sum_{t=1}^T d_t \geq \sum_{t=1}^{T_e} \min \{d_t, T - t - 1\} = \sum_{t=1}^{T_e} m_t \geq \sum_{t=1}^{T_e} m_t \geq 2^{e-1}$. Thus, $2^{\frac{e}{2}} \leq \sqrt{1 + \sum_{t=1}^T d_t} \leq \sqrt{2} \sum_{t=1}^T d_t$. Also, using $\sqrt{2} - 1 \geq 0.4$, we get

$$\mathbb{E}[\mathcal{R}(T)] \leq 5 \left( 2^{\frac{T}{2}} \sum_{t=1}^T d_t \left[ \frac{1}{2} \left\| W \right\|_F^2 + \frac{KR^2}{\gamma} + L \right] + \frac{10\sqrt{T}KR^2}{\gamma} \right)$$

$$\leq 10 \sum_{t=1}^T d_t \left[ \frac{1}{2} \left\| W \right\|_F^2 + \frac{KR^2}{\gamma} + L \right] + \frac{10\sqrt{T}KR^2}{\gamma} \right)$$

$$+ 10\sqrt{T} \left[ \frac{1}{2} \left\| W \right\|_F^2 + \frac{KR^2}{\gamma} + L \right]$$

$$\leq 10 \left[ \frac{1}{2} \left\| W \right\|_F^2 + \frac{KR^2}{\gamma} + L \right] \sqrt{T + \sum_{t=1}^T d_t}$$

We see that even for unknown $T$ and $\sum_{t \notin \mathcal{M}} d_t$, Adaptive Delaytron achieves a regret bound of $O \left( \sqrt{T + \sum_{t=1}^T d_t} \right)$.
We ran Delaytron for four different values of maximum delay. We used a pre-trained VGG-16 [18] model to extract features. Denoted by $\gamma$ value of parameter values for each dataset and delay. Figure 1 plots the final error of various real-world and synthetic data sets. The final error rate of Delaytron under different delay settings is comparable to that of other benchmarking algorithms under the 0-Delay setting. As a result, we can successfully learn a multiclass classifier under delayed bandit response settings without a priori knowledge of $T$.

We also observe that as the delay increases, Delaytron and Adaptive Delaytron take more rounds to converge. However, after sufficient trials, the error curve of Delaytron meets the error curve of Banditron. Our proposed algorithm successfully learns a multiclass classifier despite the delay in the received bandit feedback.

V. EXPERIMENTS

This section empirically evaluates the effectiveness of our proposed approaches, Delaytron and Adaptive Delaytron, on various real-world and synthetic data sets.

Datasets Used: We use CIFAR-10 [13], Fashion-MNIST [19], USPS [11], Digits, Abalone and Ecoli datasets from UCI repository [8]. To extract the features of examples in the Fashion-MNIST dataset, we use a four-layer convolutional neural network as described in [1]. Moreover, for CIFAR-10, we used a pre-trained VGG-16 [18] model to extract features.

We also performed experiments on synthetic datasets called SynSep and SynNonSep. SynSep is a 9-class, 400-dimensional artificial data set of size $10^5$. While constructing SynSep, we ensure that the dataset is linearly separable. For more details about the dataset, please refer to [12]. The idea behind SynSep is to generate a simple dataset simulating a text document. The coordinates represent different words in a small vocabulary of size 400. SynNonSep is constructed the same way as SynSep except that a 5% label noise is introduced, making the dataset non-separable.

Baselines: We compared our proposed algorithm, Delaytron, under the delayed feedback setting with the present state-of-the-art bandit algorithms (Banditron [12], and SOBA [4]) under the standard (0-Delay) setting.

Experimental Settings and Hyper-parameter Selection: We ran Delaytron for four different values of maximum delay (denoted by $D$). These are (a) 100, (b) 1000, (c) 2500 (d) 5000. We randomly sample the delay in the range $[0, D]$ during training at each trial. We ran Delaytron and other benchmarking algorithms for a wide range of the exploration parameter $\gamma$ values for each dataset and delay. Figure 1 plots the final error rates of Delaytron for the different datasets as the function of exploration parameter $\gamma$. We chose the $\gamma$ value for which the average error rate achieved is minimum. Here, the averaging is done over 20 independent runs of the algorithm.

Results: Figure 2 and Figure 3 show the plots of average error rates of Delaytron and Adaptive Delaytron (for the best value of parameter $\gamma$) against the number of instances observed so far. We plotted the result on a log-log scale to better visualize the asymptotic bounds.

By observing Figure 2 and Figure 3, we see that as the number of trials (rounds) grows, the slope of the error rate of Delaytron under different delay settings is comparable to that of other benchmarking algorithms under the 0-Delay setting.

The final error rate of Delaytron under various delays is also close to the 0-Delay approaches (e.g., Banditron and SOBA).

Figure 3 shows that the performance of Adaptive Delaytron under different delay settings is comparable to that of other benchmarking algorithms under the 0-Delay setting. As a result, we can successfully learn a multiclass classifier under delayed bandit response settings without a priori knowledge of $T$.

We also observe that as the delay increases, Delaytron and Adaptive Delaytron take more rounds to converge. However, after sufficient trials, the error curve of Delaytron meets the error curve of Banditron. Our proposed algorithm successfully learns a multiclass classifier despite the delay in the received bandit feedback.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed Delaytron algorithm that can efficiently learn multiclass classifiers with delayed bandit feedback. We show that the regret bound of the proposed approach for fixed step size is $O\left(\sqrt{\frac{2K}{\gamma} \left(\frac{T}{2} + 2 + \frac{L^2}{R^2\|W\|^2} \sum_{t=1}^{T} d_t\right)}\right)$. Here, we assumed that the loss for each missing sample is upper bounded by $L$. On the other hand, when the loss for missing samples is not upper-bounded, the regret bound of Delaytron is $O\left(\sqrt{\frac{2K}{\gamma} \left[\frac{T}{2} + 2 \sum_{t=1}^{T} d_t + |M(T)|\right]}\right)$ where $M$ is the set of missing samples in $T$ rounds. Note that the constant step size in these bounds requires the knowledge of $T$ and $\sum_{t=1}^{T} d_t$. When $T$ and $\sum_{t=1}^{T} d_t$ are unknown, we propose Adaptive Delaytron using a doubling trick for online learning and show that it achieves a regret bound of $O\left(\sqrt{T + \sum_{t=1}^{T} d_t}\right)$. Experimental results show that Delaytron and Adaptive Delaytron with delayed bandit feedback performs comparably to the state-of-the-art bandit feedback approaches, which don’t accept delays.

REFERENCES

[1] Mudit Agarwal and Naresh Manwani. Learning multiclass classifier under noisy bandit feedback. In Advances in Knowledge Discovery and Data Mining, pages 448–460, 2021.
[2] Manik Arora and Naresh Manwani. Exact passive-aggressive algorithms for multiclass classification using bandit feedbacks. In Proceedings of The 12th Asian Conference on Machine Learning, volume 129 of Proceedings of Machine Learning Research, pages 369–384, 18–20 Nov 2020.
[3] Gaurav Batra and Naresh Manwani. Multiclass classification using dilute bandit feedback. In PRICAI 2021: Trends in Artificial Intelligence, pages 63–75, 2021.
Fig. 2. Average error rate of **Delaytron** for four different values of delay ($D = 100, 1000, 2500$ and $5000$) as compared against the other benchmark algorithm under the standard (0-Delay) setting.
Fig. 3. Average error rate of Adaptive Delaytron for four different values of delay ($D = 100, 1000, 2500$ and $5000$) as compared against the other benchmark algorithm under the standard (0-Delay) setting.
Taking expectations on both sides, we get
\[
\mathbb{E}_{\mathcal{F}_t} [\|W^{t+1} - W\|_F^2] = \mathbb{E}_{\mathcal{F}_t} [\|W^t - W\|_F^2] \\
+ 2\eta \mathbb{E}_{\mathcal{F}_t} [\sum_{s \in S_t} \langle W^{t,s} - W, \hat{U}^s \rangle] + \eta^2 \mathbb{E}_{\mathcal{F}_t} [\sum_{s \in S_t} \|\hat{U}^s\|_F^2] \\
= \mathbb{E}_{\mathcal{F}_t} [\|W^t - W\|_F^2] + 2\eta \mathbb{E}_{\mathcal{F}_t} [\sum_{s \in S_t} \langle W^{t,s} - W, U^s \rangle] \\
+ \eta^2 \mathbb{E}_{\mathcal{F}_t} [\sum_{s \in S_t} \|U^s\|_F^2].
\]

Where we used the fact that
\[
\mathbb{E}_{\mathcal{F}_t} [\|W^{t,s} - W, \hat{U}^s\|_F] = \mathbb{E}_{\mathcal{F}_t} [\|W^{t,s} - W, U^s\|_F].
\]

Now, using the convexity property of the loss function \(l_H(W, (x, y))\), we know that
\[
\langle W^{t,s} - W, U^s \rangle = \langle W^t - W, U^s \rangle + \langle W^{t,s} - W^t, U^s \rangle.
\]

Using this, we get
\[
\mathbb{E}_{\mathcal{F}_t} \left[ \sum_{s \in S_t} l_H(W^s, (x_s, y_s)) \right] - \sum_{t=1}^T \sum_{s \in S_t} l_H(W, (x_s, y_s)) \\
\leq \frac{1}{2\eta} \mathbb{E}_{\mathcal{F}_t} [\|W^t - W\|_F^2 - \|W^{t+1} - W\|_F^2] \\
+ \eta^2 \sum_{s \in S_t} \|U^s\|_F^2 + \mathbb{E}_{\mathcal{F}_t} [\sum_{s \in S_t} \langle W^{t,s} - W^s, U^s \rangle].
\]

Summing from \(t = 1\) to \(T\), we get the following.
\[
\mathbb{E} \left[ \sum_{t=1}^T \sum_{s \in S_t} l_H(W^s, (x_s, y_s)) \right] - \sum_{t=1}^T \sum_{s \in S_t} l_H(W, (x_s, y_s)) \\
\leq \frac{1}{2\eta} \mathbb{E} \left[ \sum_{t=1}^T [\|W^t - W\|_F^2 - \|W^{t+1} - W\|_F^2] \\
+ \eta^2 \sum_{t=1}^T \sum_{s \in S_t} \|U^s\|_F^2 \right] + \mathbb{E} \left[ \sum_{t=1}^T \sum_{s \in S_t} \langle W^{t,s} - W^s, U^s \rangle \right].
\]

Using the fact that \(W^1 = 0 \in \mathbb{R}^{K \times d}\), we get
\[
\sum_{t=1}^T [\|W^1 - W\|_F^2 - \|W^{t+1} - W\|_F^2] = \|W^1 - W\|_F^2 \\
- \|W^{T+1} - W\|_F^2 \leq \|W^1 - W\|_F^2 = \|W\|_F^2.
\]

This gives,
\[
\mathbb{E} \left[ \sum_{t=1}^T \sum_{s \in S_t} l_H(W^s, (x_s, y_s)) \right] - \sum_{t=1}^T \sum_{s \in S_t} l_H(W, (x_s, y_s)) \\
\leq \frac{1}{2\eta} \|W\|_F^2 + \eta^2 \sum_{t=1}^T \sum_{s \in S_t} \mathbb{E}_{P_s} [\|\hat{U}^s\|_F^2] \\
+ \mathbb{E} \left[ \sum_{t=1}^T \sum_{s \in S_t} \langle W^{t,s} - W^s, U^s \rangle \right].
\]

Using Lemma 5 in [12], we know that
\[
\mathbb{E}_{P_s} [\|\hat{U}^s\|_F^2] \leq 2 \|x_s\|^2 \left( \frac{K}{\gamma} \|y_s - \hat{y}_s\| + \gamma \|y_s = \hat{y}_s\right) \leq \frac{KR^2}{\gamma}.
\]
where we used the fact that $R^2 = \max_{x \in [T]} \|x\|^2$. Using this, the bound becomes

$$E \left[ \sum_{t=1}^{T} \sum_{s \in S_t} l_H(W^s, (x_s, y_s)) \right] - \sum_{t=1}^{T} \sum_{s \in S_t} l_H(W_s, (x_s, y_s)) \leq \frac{1}{2\eta}\|W\|_F^2 + \frac{\eta K T R^2}{2\gamma} + E \left[ \sum_{t=1}^{T} \sum_{s \in S_t} (W^{t,s} - W^s, U^s) \right].$$

(2)

Each summand $(W^{t,s} - W^s, U^s)$ contributes loss proportional to the distance between the matrix $W^s$ when update $U^s$ is generated and the matrix $W^{t,s}$ when $U^s$ is applied. This distance is created by the other updates that are applied in between. Using Cauchy-Schwarz inequality, we first bound the delay terms as follows.

$$\sum_{t=1}^{T} \sum_{s \in S_t} (W^{t,s} - W^s, U^s) \leq \sum_{t=1}^{T} \sum_{s \in S_t} \|W^{t,s} - W^s\|_F \|U^s\|_F \leq R \sum_{t=1}^{T} \sum_{s \in S_t} \|W^{t,s} - W^s\|_F \leq \eta \sum_{t=1}^{T} \sum_{s \in S_t} \|U^s\|_F + \eta \sum_{t=1}^{T} \sum_{s \in S_t} \|U^s\|_F.$$

Taking expectations on both sides, we get the following bound.

$$E[\|W^{t,s} - W^s\|_F] \leq E \left[ \sum_{t=1}^{T} \sum_{s \in S_t} \|U^s\|_F + \sum_{r=s}^{t-1} \sum_{s \in S_t} \|U^s\|_F \right].$$

Using the concavity property of the square root function, we observe that $E_{P_s}[\|U^s\|_F] = E_{P_s}[\sqrt{\|U^s\|_F^2}] \leq \sqrt{E_{P_s}[\|U^s\|_F^2]} \leq \sqrt{K R^2 \gamma}$. Thus,

$$E[\|W^{t,s} - W^s\|_F] \leq \sqrt{\frac{K R^2 \gamma}{\gamma}} \left( |S_{t,s}| + \sum_{r=s}^{t-1} |S_r| \right) \leq \sqrt{\frac{K R^2 \gamma}{\gamma}} \left( |S_{t,s}| + \sum_{r=s}^{t-1} |S_r| \right)$$

(4)

Using eq.(4) and (3) in eq.(2), we get the following.

$$\sum_{t=1}^{T} \sum_{s \in S_t} l_H(W^s, (x_s, y_s)) - \sum_{t=1}^{T} \sum_{s \in S_t} l_H(W_s, (x_s, y_s)) \leq \frac{1}{2\eta}\|W\|_F^2 + \frac{K R^2}{\gamma} \left[ \frac{\eta T}{2} \sum_{t=1}^{T} \sum_{s \in S_t} \left( |S_{t,s}| + \sum_{r=s}^{t-1} |S_r| \right) \right].$$

(5)

Using Lemma 1, we have $\sum_{t=1}^{T} \sum_{s \in S_t} |S_{t,s}| + \sum_{r=s}^{t-1} |S_r| \leq 2 \sum_{t \in \mathcal{M}} d_t$. Using this in eq.(5), we get

$$E \left[ \sum_{t=1}^{T} \sum_{s \in S_t} l_H(W^s, (x_s, y_s)) \right] - \sum_{t=1}^{T} \sum_{s \in S_t} l_H(W_s, (x_s, y_s)) \leq \frac{1}{2\eta}\|W\|_F^2 + \frac{\eta K R^2}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t \right].$$

(6)

**APPENDIX B**

**PROOF OF THEOREM 2**

**Proof.**

$$E \left[ \sum_{t \in \mathcal{M}} l_H(W^t, (x_t, y_t)) - \sum_{t \in \mathcal{M}} l_H(W_t, (x_t, y_t)) \right] \leq E \left[ \sum_{t \in \mathcal{M}} l_H(W^t, (x_t, y_t)) \right] \leq L|\mathcal{M}|$$

(7)

We get the regret as follows using Theorem 1 and (7).

$$R(T) \leq \frac{1}{2\eta}\|W\|_F^2 + \frac{\eta K R^2}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t \right] + L|\mathcal{M}|$$

Using $\eta = \sqrt{\frac{2K R^2 \gamma}{T + 2 \sum_{t \in \mathcal{M}} d_t}}$, we get the following bound.

$$R(T) \leq R\|W\|_F \sqrt{\frac{2K}{\sqrt{\gamma}} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t \right] + L|\mathcal{M}|}$$

**APPENDIX C**

**PROOF OF COROLLARY 1**

**Proof.** Thus,

$$R(T) \leq R\|W\|_F \sqrt{\frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t \right] + 1 \frac{K L^2 m (m+1)}{2 \gamma}}$$

Using the concavity property of the square root function, we observe that $E_{P_s}[\|U^s\|_F^2] = E_{P_s}[\sqrt{\|U^s\|_F^2}] \leq \sqrt{E_{P_s}[\|U^s\|_F^2]} \leq \sqrt{K R^2 \gamma}$. Thus,

$$E[\|W^{t,s} - W^s\|_F] \leq \sqrt{\frac{K R^2 \gamma}{\gamma}} \left( |S_{t,s}| + \sum_{r=s}^{t-1} |S_r| \right)$$

(4)

Using eq.(4) and (3) in eq.(2), we get the following.

$$E \left[ \sum_{t=1}^{T} \sum_{s \in S_t} l_H(W^s, (x_s, y_s)) \right] - \sum_{t=1}^{T} \sum_{s \in S_t} l_H(W_s, (x_s, y_s)) \leq \frac{1}{2\eta}\|W\|_F^2 + \frac{K R^2}{\gamma} \left[ \frac{\eta T}{2} \sum_{t=1}^{T} \sum_{s \in S_t} \left( |S_{t,s}| + \sum_{r=s}^{t-1} |S_r| \right) \right].$$

(5)

The second inequality follows from the concavity property of $f(x) = \sqrt{x}$. 

□
We get the overall regret as follows using Theorem 1 and eq.(8).

\[ \mathcal{R}(T) \leq |\mathcal{M}| \mathcal{R}(\|W\|_F + \frac{\eta KTR}{\gamma}) \]

We get the overall regret as follows using Theorem 1 and eq.(8).

\[ \mathcal{R}(T) \leq |\mathcal{M}| \mathcal{R}(\|W\|_F + \frac{1}{2\eta} \|W\|_F^2 + \frac{\eta K R^2}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}|T \right] \]

Using \( \eta = \sqrt{\frac{2K R^2}{T^2 + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}|T}} \), we get the following bound.

\[ \mathcal{R}(T) \leq R \|W\|_F \left[ |\mathcal{M}| + \sqrt{\frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}|T \right]} \right] \]

\[ \text{APPENDIX D} \]

\textbf{PROOF OF THEOREM 3}

Proof.

\[
\mathbb{E} \left[ \sum_{t \in \mathcal{M}} \left\{ l_H(W^t, (x_t, y_t)) - l_H(W, (x_t, y_t)) \right\} \right]
\leq \mathbb{E} \left[ \sum_{t \in \mathcal{M}} \left\langle \nabla l_H(W^t, (x_t, y_t)), W^t - W^* \right\rangle \right]
\leq \mathbb{E} \left[ \sum_{t \in \mathcal{M}} \| \nabla l_H(W^t, (x_t, y_t)) \|_F \cdot \| W^t - W^* \|_F \right]
\]

But, \( \| \nabla l_H(W^t, (x_t, y_t)) \|_F \leq \| x_t \|_2 \leq R \). Also, \( \| W^t - W \|_F \leq \| W^t \|_F + \| W \|_F \). But, \( W^t = \eta \sum_{t=1}^{T-1} \sum_{r \in S_r} \tilde{U}_r^t \). Thus, \( \| W^t \|_F \leq \eta \sum_{t=1}^{T-1} \sum_{r \in S_r} \| \tilde{U}_r^t \|_F \). Taking expectation, we get, \( \mathbb{E}[\| W^t \|_F] \leq \eta \sum_{t=1}^{T-1} \sum_{r \in S_r} \mathbb{E}[\| \tilde{U}_r^t \|_F] \leq \eta R \sqrt{\frac{K}{\gamma} \sum_{t=1}^{T-1} |S_r|}. \sum_{t=1}^{T-1} |S_r| \) is the number of feedbacks received till time \( t - 1 \) which is not more than \( t - 1 \). Thus, \( \mathbb{E}[\| W^t \|_F] \leq \frac{\eta R \sqrt{K T R}}{\gamma} (t - 1) \leq \frac{2K R}{\gamma} t \). So, we get

\[
\mathbb{E} \left[ \sum_{t \in \mathcal{M}} l_H(W^t, (x_t, y_t)) \right] - \mathbb{E} \left[ \sum_{t \in \mathcal{M}} l_H(W, (x_t, y_t)) \right]
\leq |\mathcal{M}| \mathcal{R}(\|W\|_F + \frac{\eta KTR}{\gamma}) \tag{8}
\]

\[ \text{APPENDIX E} \]

\textbf{PROOF OF COROLLARY 2}

Proof. We see that

\[
\sqrt{\frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}|T \right] + |\mathcal{M}|}
\leq \frac{1}{2} \sqrt{\frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + |\mathcal{M}|T \right] + \frac{K m(m + 1)}{\gamma}}
\leq \sqrt{\frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t \in \mathcal{M}} d_t + m(m + 1)/2 + |\mathcal{M}|T \right]}
\leq \sqrt{\frac{2K}{\gamma} \left[ \frac{T}{2} + 2 \sum_{t=1}^{T} d_t + |\mathcal{M}|T \right]}. \]