On the Importance of Randomization in Discriminative Feature Feedback

Valentio Iverson VIVERSON@UWATERLOO.CA
Cheriton School of Computer Science, University of Waterloo

Tosca Lechner TOSCA.LECHNER@VECTORINSTITUTE.AI
Vector Institute for Artificial Intelligence, Toronto, Canada

Sivan Sabato SABATOS@MCMASTER.CA
Department of Computing and Software, McMaster University
Canada CIFAR AI Chair, Vector Institute for Artificial Intelligence, Toronto, Canada
Faculty of Computer and Information Science, Ben-Gurion University of the Negev

Editors: Steve Hanneke and Tor Lattimore

Abstract

Discriminative Feature Feedback (DFF) (Dasgupta et al., 2018) is an interactive learning protocol in which a learner attempts to predict labels based on online feedback about previous correct labels, as well as *discriminative features*. Recent work (Bar Oz et al., 2025) studied DFF learning with general teacher classes and defined a dimension that characterizes the optimal mistake bound for deterministic algorithms in the realizable setting. In this work, we show that in sharp contrast to Online Learning, in DFF there can be an unbounded ratio between the optimal mistake bound of deterministic algorithms and that of randomized algorithms, even in the realizable setting. We further show that in this case, also non-realizable learning can have a mistake bound that does not depend on the dimension at all. This result relies on a new algorithmic technique that allows introducing new candidate hypotheses incrementally and could be of independent interest. We further show that in DFF, there can be a significant difference between the obtainable mistake bounds against an oblivious adversary and against an adaptive adversary, again in contrast to Online Learning. Our work shows that once richer feedback than labels is allowed, the landscape of randomized versus deterministic algorithms becomes significantly more involved, and raises new questions on characterizing the optimal mistake bound under differing randomization regimes.

Keywords: Discriminative Feature Feedback, Online Learning, randomization, adaptivity

1. Introduction

Discriminative Feature Feedback (Dasgupta et al., 2018) is an interactive learning protocol in which a learner attempts to predict labels based on online feedback about previous correct labels, as well as *discriminative features*. A discriminative feature is a property of an example that explains why its label is different from the label of another example, which the learner assumed has the same label. This protocol formalizes a certain type of rich feedback that is common in human learning. For instance, a user may indicate that they don't like a certain movie, even though they liked the previous movie of the same director, because they don't like the main actor, who is different in this new movie then in the previous movie.

Recent work (Bar Oz et al., 2025) studied the theoretical properties of Discriminative Feature Feedback (DFF) with a general *teacher class*. Here, each *teacher* defines both the labeling function and the type of feature feedback that would be provided in each possible interaction with the learner. In that work, a *DFF Dimension* was defined, and it was shown that this dimension is equal to the optimal mistake bound of any teacher class under the DFF protocol in the realizable setting, assuming a deterministic algorithm. However, in the non-realizable setting, it was shown that some teacher classes cannot have sub-linear regret even when the DFF dimension is finite, at least under some randomization models. In contrast, some other teacher classes are equivalent to Online Learning (under the 0-1 loss, which we assume henceforth), and thus the Weighted Majority algorithm (Littlestone and Warmuth, 1994) obtains sub-linear regret for those classes. This result revealed that unlike Online Learning, in which the Littlestone dimension fully characterizes the optimal obtainable regret for both the realizable and the non-realizable setting, the DFF Dimension does not provide the same characterization power. It was left open whether there are teacher classes in which the regret in the non-realizable setting can have an even smaller dependence in the dimension than the dependence in Online Learning.

In this work, we answer this question in the affirmative. Moreover, our proof provides a crucial observation: unlike Online Learning, in which optimal mistake bounds for deterministic and randomized algorithms are at most a factor of two apart[1] (Ben-David et al., 2009), in DFF, the ratio between the mistake bound of an optimal deterministic algorithm and that of an optimal randomized algorithm can be infinite. We show a teacher class in which the DFF dimension is infinite, yet a randomized algorithm exists that makes a constant number of mistakes in the realizable setting. Moreover, in the non-realizable setting, we provide for this teacher class a new algorithm that obtains vanishing regret, which we also show to be near optimal. This algorithm combines the Weighted Majority algorithm (Littlestone and Warmuth, 1994) with random exploration, and further incorporates a new technique which could be of independent interest: instead of initializing with the full set of possible teachers, it initializes with a single teacher, and adds new teachers if evidence of their plausibility is provided. This allows the regret bound to depend on the number of introduced teachers instead of the number of a-priori possible teachers.

The nuanced importance of randomization in DFF is further demonstrated by comparing *oblivious* and *adaptive* adversaries of randomized algorithms. Oblivious adversaries make their decisions before observing any of the algorithm's outputs. Adaptive adversaries can make decisions round by round, using the outputs of the previous rounds as input for the decision of the next round. In Online Learning, the Weighted Majority algorithm is a randomized algorithm that is optimal for both oblivious and adaptive adversaries and obtains the same vanishing regret for both. In contrast, for DFF, we show a strong separation between oblivious and adaptive adversaries, in that an adaptive adversary can prevent vanishing regret while in the same teacher class, vanishing regret can be obtained for oblivious adversaries.

Our technical contributions include, for the realizable setting:

- A teacher class $\mathcal{T}_{[0,1]}$ with infinitely many labels that has an infinite DFF dimension, but has a randomized algorithm that makes a single mistake with probability 1 against both an oblivious and an adaptive adversary.

1. The factor of two gap easily extends to multiclass settings, although this was not explicitly discussed in (Ben-David et al., 2009).

- A teacher class, denoted $\mathcal{T}_3$, with 3 labels, that has an infinite DFF dimension but has a randomized algorithm that makes a constant number of mistakes in expectation against both an oblivious and an adaptive adversary.

Our technical contributions for the non-realizable setting include:

- A mistake lower bound that holds for any teacher class against oblivious and adaptive adversaries.
- A randomized algorithm for $\mathcal{T}_3$ that obtains vanishing regret against an oblivious adversary, and is near optimal with respect to the lower bound above. The algorithm uses a technique of increasing the hypothesis class incrementally, which could be of independent interest.
- A lower bound showing that vanishing regret is impossible for $\mathcal{T}_3$ against an adaptive adversary, showing a separation between adaptive and oblivious adversaries.
- An algorithm for $\mathcal{T}_{[0,1]}$ that obtains linear regret against an adaptive adversary.

These results demonstrate the importance of randomization properties in settings with rich feedback that go beyond the classical Online Learning protocol. Moreover, they open new questions regarding the existence of a "randomized" DFF dimension and the possibility that it can provide a tighter characterization of optimal mistake bounds for DFF.

Related work

Various types of learning protocols with rich feedback have been used in applications, (see, e.g., Mosqueira-Rey et al., 2022), including machine-generated explanations (Teso and Kersting, 2019) with human corrections, learning with explanations in neural networks (Schramowski et al., 2020) and explanations for Large Language Models (Lampinen et al., 2022). Auditing with explanations was studied in Yadav et al. (2024). A theoretical perspective on interactive learning has been provided in Hanneke et al. (2022), in which any binary-valued query is allowed.

Feature feedback to improve learning outcomes has been suggested as early as Croft and Das (1990), and variations of that idea have been used in Raghavan et al. (2005); Druck et al. (2008); Settles (2011); Mac Aodha et al. (2018). The DFF protocol was first formally defined in Dasgupta et al. (2018), and the non-realizable setting was first studied in Dasgupta and Sabato (2020). Mistake bounds for specific teacher classes were derived in Dasgupta et al. (2018) for the realizable setting and in Dasgupta and Sabato (2020); Sabato (2023) for the non-realizable setting. Bar Oz et al. (2025) studied DFF with general teacher classes, providing a combinatorial definition of the DFF dimension, which gives the optimal mistake bound of a given teacher class when using a deterministic algorithm. That work further provided a mistake upper bound for the non-realizable setting, using a deterministic algorithm. This upper bound was shown to be tight for a specific teacher class, for randomized algorithms with an adaptive adversary. However, whether this upper bound is tight also for oblivious adversaries remains open.

In Online Learning, the optimal expected mistake bound of a randomized algorithm with an oblivious adversary is at least half the optimal mistake bound of a deterministic algorithm (Ben-David et al., 2009). The Littlestone dimension of a hypothesis class gives the optimal mistake bound for deterministic algorithms in the realizable setting (Littlestone, 1988). Filmus et al. (2023) define the Randomized Littlestone dimension that gives the optimal mistake bound for a randomized

algorithm in the realizable setting. The value of the Randomized Littlestone dimension is at least half of the Littlestone dimension.

A deterministic algorithm in the non-realizable setting cannot have vanishing regret, since its mistake bound is at least twice the minimal number of label errors. Nonetheless, the optimal mistake bound of a deterministic algorithm is still at most twice that of the optimal randomized algorithm.

Adaptive adversaries have been shown to be strongly separated from oblivious adversaries in a more general setting of online prediction with bandit feedback and general losses (Arora et al., 2012). In this setting, vanishing regret can be obtained for oblivious adversaries, while an adaptive adversary can force a constant loss at almost every prediction round. However, this separation employs losses that depend on all previous predictions, and not only the prediction of the current round. Thus, it does not apply to Online Learning or to Discriminative Feature Feedback.

2. Preliminaries

The DFF learning protocol with a general teacher class (Dasgupta et al., 2018; Bar Oz et al., 2025) is defined as follows. There is a domain of examples $\mathcal{X}$ and a domain of labels $\mathcal{Y}$. There is a set $\Phi \subseteq \{0,1\}^{\mathcal{X}}$ of Boolean features of domain examples. $\perp$ is a special element not in Φ. In every round of the DFF protocol:

- An example x_t is presented to the learner;
- The learner provides a prediction $\hat{y}_t$, and an example $\hat{x}_t$ that was previously observed with that label. This example is the explanation for the predicted label: "x_t is predicted to have label $\hat{y}_t$ because $\hat{x}_t$ was labeled $\hat{y}_t$".
- If the prediction is incorrect, the environment provides the correct label y_t of x_t and a feature $\phi_t \in \Phi$ that explains why x_t does not have the same label as $\hat{x}_t$. ϕ_t is satisfied by x_t but not by $\hat{x}_t$.

Since the algorithm requires previous examples to predict, it is initialized with a limited set of given labeled examples, which is called the *history*, and denoted $H \subseteq \mathcal{X} \times \mathcal{Y}$. We say the algorithm makes a mistake in round t, if $\hat{y}_t \neq y_t$. The goal of the algorithm is to make a small number of mistakes during its run.

A *teacher* specifies the possible feedback that can be provided to the learner.

Definition 1 (DFF Teacher, Bar Oz et al. 2025) *A* teacher $\mathcal{T}$ *over* $\mathcal{X}, \mathcal{Y}, \Phi$ *is a pair* (f, ψ)*, where* $f : \mathcal{X} \to \mathcal{Y}$ *is a labeling function and* $\psi : \mathcal{X} \times \mathcal{X} \to \Phi \cup \{\perp\}$ *is a feature feedback function, such that for all* $x, \hat{x} \in \mathcal{X}$*, if* $f(x) \neq f(\hat{x})$ *then* $\phi := \psi(x, \hat{x}) \in \Phi$*. In addition,* ϕ *satisfies* x *and does not satisfy* $\hat{x}$*.*

A *teacher class* $\mathcal{T}$ over $\mathcal{X}, \mathcal{Y}, \Phi$ is a set of teachers over $\mathcal{X}, \mathcal{Y}, \Phi$. We usually assume fixed $\mathcal{X}, \mathcal{Y}, \Phi$, and omit the expression "over $\mathcal{X}, \mathcal{Y}, \Phi$" when clear from context.

We say a that a protocol sequence is realizable in $\mathcal{T}$, if there is a teacher $(f, \psi) \in \mathcal{T}$, such that for every round t, we have $f(x_t) = y_t$ and $\phi_t = \psi(x_t, \hat{x}_t)$. Bar Oz et al. (2025) define the DFF dimension of a pair of a teacher class and history, and show that it is equal to the optimal mistake bound of a deterministic algorithm in the realizable setting.

Setting	realizable (best class)	realizable (worst class)	k-non-realizable (best class)	k-non-realizable (worst class)
deterministic	d [B25]	d [B25]	at least $2k+d$ [O8]; $2k + O(\sqrt{kd} + d)$ [F23]	$\Theta(kd)$ [B25]
randomized (adaptive adv.)	$\Theta(1)$ [T2]	d [F23,B25]	$k+\Omega(\sqrt{k})$ [O7]; $\min(2k+\widetilde{O}(\sqrt{k}), k+O(\sqrt{kd}+d)$ [T14,B09]	$\Theta(kd)$ [B25]
randomized (oblivious adv.)	$\Theta(1)$ [T2]	d [F23,B25]	$k+\widetilde{\Theta}(\sqrt{k})$ [T9,O7]	$k+\Omega(\sqrt{kd}+d)$ [B09]; $O(kd)$ [B25]

Table 1: In this table, we summarize the mistake bounds in the settings that we study, as a function of the DFF-dimension d (assuming $d \geq 1$) and of k in the k-non-realizable setting. We also distinguish between learning bounds for the best and the worst class in a given setting, since in many cases the dependence of the optimal mistake bounds on d, k differ between classes. Our new results are referred to as T2: Theorem 2, T9: Theorem 9, T14: Theorem 14, O7: Observation 7, O8: Observation 8. The other results exist in or follow directly from the literature: F23: Filmus et al. (2023), B25: Bar Oz et al. (2025), B09: Ben-David et al. (2009).

We say a sequence is k-non-realizable if there is a teacher $(f, \psi) \in \mathcal{T}$, such that in all but at most k rounds, we have $f(x_t) = y_t$ and $\phi_t = \psi(x_t, \hat{x}_t)$. We call such a teacher the "true teacher" and the set of up to k rounds not consistent with (f, ψ) "exception rounds" with respect to (f, ψ).

While we will usually discuss mistake bounds, another standard measure is the regret, which is defined as the difference between the expected number of mistakes and k. We say the algorithm has vanishing (or sublinear) regret if its regret is $o(k)$.

When discussing randomized learners, it is crucial to define which class of adversaries is considered. In this work, we consider two types of adversaries. The *oblivious adversary* fixes in advance a sequence of instances $x_1, \dots x_T$ and a sequence of teachers $(f_1, \psi_1), \dots, (f_T, \psi_T)$. In every round, the algorithm outputs $\hat{y}_t, \hat{x}_t$ and the feedback of the adversary is $y_t = f_t(x_t)$ and $\phi_t = \psi_t(x_t, \hat{x}_t)$. For a given algorithm $\mathcal{A}$, a teacher class $\mathcal{T}$ and a history H, we denote the worst-case expected number of mistakes with an oblivious adversary in the k-non-realizable setting as $\mathcal{M}_k^{\text{obl}}(\mathcal{A}, \mathcal{T}, H)$.

In contrast, the *adaptive adversary* fixes the instance x_t and teacher (f_t, ψ_t) only after having observed the responses of the algorithm of previous rounds $(\hat{y}_1, \hat{x}_1), \dots (\hat{y}_{t-1}, \hat{x}_{t-1})$. We can think of the adaptive adversary as being allowed to choose the "true" teacher as well as the exception rounds in retrospect, while the oblivious adversary needs to commit to these choices in advance.

For a given algorithm $\mathcal{A}$, a teacher class $\mathcal{T}$ and a history H, we denote the worst-case expected number of mistakes with an adaptive adversary in the k-non-realizable setting as $\mathcal{M}_k^{\text{adapt}}(\mathcal{A}, \mathcal{T}, H)$. Trivially, $\mathcal{M}_k^{\text{obl}}(\mathcal{A}, \mathcal{T}, H) \leq \mathcal{M}_k^{\text{adapt}}(\mathcal{A}, \mathcal{T}, H)$. We omit k to indicate the realizable setting.

3. Main Results

In this paper, we show that the gap between randomized learners and deterministic learners in the DFF setting can be arbitrarily large. In Table 1, we summarize the landscape of learning bounds in relation to the DFF-dimension d of Bar Oz et al. (2025) and the k non-realizability parameter, including our new results. Some of the existing bounds are direct implications of previous bounds for Online Learning, since it can be seen as a special case of DFF (Bar Oz et al., 2025).

In Section 4, we show this separation for the realizable case: Theorem 2 shows that in the realizable case, there are classes for which deterministic learners make an arbitrary number of mistakes, that is $d = \infty$, but which can be learned by randomized learners with a constant expected number of mistakes. In Section 5, we show a separation between deterministic and randomized learners in the non-realizable case. In particular, we show that there are classes with $d = \infty$ that can be learned with vanishing regret in the case of oblivious adversaries (Theorem 9), where we obtain a mistake bound of $k + \widetilde{O}(\sqrt{k})$ using a new algorithm, and show that this is tight up to log factors.

For adaptive adversaries, we show that there are classes with $d = \infty$ and a mistake bound $2k + \widetilde{O}(\sqrt{k})$ using a randomized algorithm (Theorem 14). While this result does not yield vanishing regret, it does show a separation between randomized learners with adaptive adversaries and deterministic adversaries. Lastly, we show a separation between randomized learning with adaptive and oblivious adversaries, showing that there are classes than can be learned with vanishing regret in the oblivious case, but not in the adaptive case (Theorem 15).

This difference between adaptive and oblivious cases stands in contrast to Online Learning, where the two adversary models have essentially same optimal mistake bounds, as we discuss in Section 1. Since Online Learning is a special case of DFF learning, this means that the same holds for DFF for some teacher classes. However, our results show that in general, the relationship between adaptive and oblivious adversaries in DFF is class-dependent: for some classes, there is a separation between oblivious and adaptive adversaries that is similar to the separation between randomized and deterministic learners in the Online Learning, while for other classes there is no such separation.

4. Randomized Learners for DFF in the realizable setting

In this section, we show that even in the realizable case, there is a strong separation between the power of randomized and deterministic learners in DFF. We give two examples that demonstrate the separation. The first example provides a crisper separation, as it admits a randomized algorithm that makes at most a single mistake for adaptive and oblivious adversaries, while any deterministic algorithm makes a mistake in every round. However, this example uses an infinite history and uncountably many labels. The second example provides a constant mistake bound in expectation for a randomized algorithm, and has an unbounded number of mistakes for any deterministic algorithm. This example uses only three labels and a history with three examples. The results are summarized in the following theorem.

Theorem 2 *There exists a class $\mathcal{T}$ with history H, such that such that there exists a randomized algorithm which makes at most a single mistake with probability* 1*. On the other hand, for any deterministic algorithm there is an adversary that causes a mistake in every round on an infinite sequence. In addition, there exists a class $\mathcal{T}$ with history H with the same lower bound for deterministic algorithms, that has 3 labels and a finite history, and satisfies $\mathcal{M}^{\text{adapt}}(\mathcal{A}, \mathcal{T}, H) \leq 2$.*

This result on DFF learning stands in contrast to Online Learning, where the expected number of mistakes of a randomized Online Learning algorithm is always at least half of the mistake bound for the best deterministic algorithms (Ben-David et al., 2009). We prove the theorem by studying the examples below.

4.1. An example with an uncountable history and label set

For the first example, consider a case with an uncountable number of labels, $\mathcal{Y} = [0, 1]$ and assume the domain $\mathcal{X} = \mathbb{R}$. Let $g(x) := -x \cdot 1_{x\in[-1,0]}$. Consider following the class of labeling functions:

$$\mathcal{F} = \{f : \mathbb{R} \to [0, 1] \mid \forall x \notin [0, 1], f(x) = g(x)\}.$$

We define the feature set $\Phi = \{1_x \mid x \in \mathbb{R}\} \cup \{\phi_{f,x} \mid f \in \mathcal{F}, x \in [0, 1]\}$. Features of the form $\phi_{f,x}$ are constructed to identify the labeling function exactly, as follows: Given $z \in \mathbb{R}$, if $z \leq 1$, then $\phi_{f,x}(z) = 1_{z=x}$. If $z > 1$, then decompose $z = a(z) + b(z)$, where $a \in \mathbb{Z}$ and $b \in [0, 1)$. Define $\phi_f(z)$ to be the $|a(z)|$'th bit in the binary representation of $f(b(z))$. It is easy to verify that the function $\phi_{f,x}$ fully determines f. We define a teacher class such that that in most cases, the feature feedback of the teacher identifies the labeling function exactly, but in some cases the feedback is completely uninformative. To be precise, for $f \in \mathcal{F}$, define ψ_f as follows:

$$\psi_f(x, \hat{x}) = \begin{cases} \perp & \text{if } f(x) = f(\hat{x}) \\ 1_x & \text{if } f(\hat{x}) = (f(x) + \frac{1}{2}) \bmod 1 \\ \phi_{f,x} & \text{if } x \in [0, 1] \text{ and } \hat{x} \in [-1, 1] \text{ and } f(\hat{x}) \notin \{f(x), (f(x) + \frac{1}{2}) \bmod 1\} \\ 1_x & \text{otherwise} \end{cases}$$

Note that the returned feature always satisfies $\psi_f(x, \hat{x})(x) = 1$ and $\psi_f(x, \hat{x})(\hat{x}) = 0$ for $\hat{x} \neq x$. We now define the class $\mathcal{T}_{[0,1]} = \{(f, \psi_f) : f \in \mathcal{F}\}$ and the history $H_{[0,1]} = \{(-x, x) : x \in [0, 1]\}$. Lemmas 4 and 3 below prove Theorem 2.

Lemma 3 *Every deterministic algorithm for $(\mathcal{T}_{[0,1]}, H_{[0,1]})$ in the realizable setting makes a mistake in every round in the presence of a worst-case adversary.*

Proof The idea is to make sure that the adversary always returns feature feedbacks of the form $\mathbf{1}_x$, i.e., un-informative feature feedback. Set $x_t = \frac{1}{t+2}$. Thus, for any $t \neq t'$, we have $x_t \neq x_{t'}$. For every sequence $S = (x_t, y_t)_{t\in\mathbb{N}}$ there exists some $f \in \mathcal{F}$ that is consistent with S. Furthermore, for an algorithm's response $(\hat{x}_t, \hat{y}_t)$, the adversary will respond with the label $y_t = (\hat{y}_t - \frac{1}{2}) \bmod 1$ and the feature feedback $\mathbf{1}_x$.

For any function $f \in \mathcal{F}$ consistent with (x_t, y_t), we have $f(\hat{x}_t) = \hat{y}_t = (y_t + \frac{1}{2}) \bmod 1 = (f(x_t) + \frac{1}{2}) \bmod 1$. Thus, the feature feedback $\mathbf{1}_x$ is consistent with the feature feedback function $\psi_f(x_t, \hat{x}_t)$. Since this holds for every element of the sequence, S is consistent with the teacher (f, ψ_f) whenever f is consistent with $\{(x_t, y_t)\}_{t=1}^T$. Thus, the adversary's responses are consistent with a teacher in the class $\mathcal{T}_{[0,1]}$. It is also clear that the adversary presented here forces a mistake in every round. ■

$(\mathcal{T}_{[0,1]}, H_{[0,1]})$ is thus not learnable with a finite mistake bound in the realizable case by any deterministic learner. Furthermore, a worst-case adversary can force a mistake for an infinite sequence, thus the class is not even learnable in the sense of non-uniform learnability (see Bousquet

et al., 2021). We now show that in contrast, a randomized algorithm for this class can achieve a mistake bound of 1.

Lemma 4 *The pair $(\mathcal{T}_{[0,1]}, H_{[0,1]})$ has an optimal expected mistake bound of 1 using randomized algorithms against both oblivious and adaptive adversaries. Specifically, there exists a learner that for every sequence makes at most 1 mistake with probability 1.*

Proof First, we note that if $x \neq [0,1]$, then $f(x) = g(x)$. Thus, on elements $x \notin [0,1]$ the algorithm can just return the correct label $g(x)$. Thus, without loss of generality we can assume $x_t \in [0,1]$. The true label y_t is fixed by the adversary. The algorithm draws a label $\hat{y}_t \in [0,1]$ uniformly at random and outputs $(-\hat{y}_t, \hat{y}_t) \in H_{[0,1]}$. With probability 1, $y_t \neq \hat{y}_t \neq (y_t + \frac{1}{2}) \bmod 1$. Thus, every consistent feature feedback function must be of the type $\phi_{f,x}$. Since ϕ_f determines f, in every subsequent round t, for the example x_t the algorithm will output $(-f(x_t), f(x_t)) \in H_{[0,1]}$, which gives a correct label. Thus, the algorithm only makes one mistake. ■

4.2. An example with 3 labels and a finite history

Consider the domain $\mathcal{X} = \mathbb{N}$ and the label space $\mathcal{Y} = \{0,1,2\}$. We denote the class of all possible labeling functions $\mathcal{F}_3 = \{f : \mathbb{N} \to \mathcal{Y}\}$. We define the feature set $\Phi = \{1_x \mid x \in \mathbb{N}\} \cup \{\phi_{f,x,\hat{x}} \mid f \in \mathcal{F}_3, x, x' \in \mathbb{N}\}$. Features of the form $\phi_{f,x,\hat{x}}$ are constructed to identify the labeling function exactly, as follows:

$$\phi_{f,x,\hat{x}}(z) = \begin{cases} 1 & \text{if } z = x \\ 0 & \text{if } z \neq x \text{ and } z \leq 3\max(x,\hat{x}) \\ 1 & \text{if } z \geq 3\max(x,\hat{x}) \text{ and } f(\lfloor \frac{z}{3} \rfloor - \max(x,\hat{x})) \equiv z \pmod 3 \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $\phi_{f,x,\hat{x}}$ determines f and satisfies $\phi_{f,x,\hat{x}}(x) = 1$ and $\phi_{f,x,\hat{x}}(\hat{x}) = 0$.

For every $f \in \mathcal{F}_3$, we define a corresponding feature feedback function ψ_f as follows:

$$\psi_f(x,\hat{x}) = \begin{cases} \perp & \text{if } f(x) = f(\hat{x}) \\ 1_x & \text{if } f(x) - f(\hat{x}) \equiv 1 \pmod 3, \\ \phi_{f,x,\hat{x}} & \text{if } f(x) - f(\hat{x}) \equiv 2 \pmod 3. \end{cases}$$

We now consider the teacher classes $\mathcal{T}_3 = \{(f, \psi_f) : f \in \mathcal{F}_\infty\}$ with the history $H_3 := \{(1,1),(2,2),(3,0)\}$. We first show that a deterministic learner can be forced to make an unbounded number of mistakes. Lemmas 5 and 6 below prove the second part of Theorem 2.

Lemma 5 *Every deterministic algorithm for $(\mathcal{T}_3, H_3)$ makes an unbounded number of mistakes in the presence of a worst-case adversary if there are sufficiently many rounds.*

Proof The adversary sets $x_t = 3 + t$ for every $t \in \mathbb{N}$. For a given x_t, if the algorithm outputs $(\hat{y}_t, \hat{x}_t)$, the adversary sets the label $y_t = \hat{y}_t + 1 \pmod 3$ and the feature feedback $\phi_t = \mathbf{1}_x$. Since $\mathcal{F}_\infty$ includes all possible labelings of $\mathbb{N}$, there is a function $f \in \mathcal{F}_\infty$ with $f(x_t) = y_t$ for all $t \in \mathbb{N}$. Furthermore, since for every round t, $y_t - \hat{y}_t \equiv 1 \pmod 3$, and we have $\psi_f(x_t, \hat{x}_t) = \mathbf{1}_x$. Thus,

the adversary's feedback is consistent with some teacher $(f, \psi_f) \in \mathcal{T}_3$, and the algorithm makes a mistake in every round. ■

At the same time, the following lemma shows a randomized learner that makes a constant expected number of mistakes for oblivious and adaptive adversaries.

Lemma 6 *There exists a learner $\mathcal{A}$ with expected mistake bound $\mathcal{M}^{\text{adapt}}(\mathcal{A}, \mathcal{T}_3, H_3) \leq 2$.*

Proof We define the learner $\mathcal{A}$ as follows: This learner predicts a label uniformly at random from $\{0, 1, 2\}$ for any x_t, unless it can infer the correct label of x_t, which occurs if x_t is in the history or was seen before, or if at a previous round f was revealed via a feature feedback of the form $\phi_{f,x,\hat{x}}$. Every random label prediction leads to a mistake with probability $2/3$, and given that it is a mistake it leads to such feature feedback with probability $1/2$. Thus, the expected number of mistakes can be upper bounded by the expected number of random draws it takes to observe a revealing feature out of mistake rounds, which is equal to 2. ■

5. Randomized Learners for DFF in the k-non-realizable setting

In this section, we show several separation results for randomized learners in the k-non-realizable setting. First, we observe that a lower bound for randomized algorithms can be easily derived from known results for Online Learning.

Observation 7 *For any non-trivial teacher class, every randomized learning algorithm for DFF in the k-non-realizable setting has a worst-case expected mistake bound of at least $k + \Omega(\sqrt{k})$, with respect to both adaptive and oblivious adversaries.*

This lower bound follows directly from applying a part of the analysis of Ben-David et al. (2009, Lemma 14), where a lower bound is provided for Online Learning, which is equal to $T/2 + \Omega(\sqrt{T})$ for classes with Littlestone dimension of 1, when running for T rounds. This lower bound relies on the following observation: if a balanced coin is tossed T times, then the expected number of appearances of the less frequent result is $T/2 - \Theta(\sqrt{T})$. On the other hand, the expected number of times a prediction of the coin toss is successful, for any algorithm, is $T/2$.

For DFF, if there is at least one example which is labeled differently by two different teachers, the adversary can keep presenting that same example for T rounds, randomly choosing one of the two teachers in each round. Suppose for contradiction that some algorithm has a mistake upper bound of $k + o(\sqrt{k})$ for a k-non-realizable run. Then, since the expected value of k in these runs is $T/2 - \Theta(\sqrt{T})$, this gives an expected mistake bound of $o(T/2)$, a contradiction. We conclude that any algorithm for DFF in the k-non-realizable setting has a mistake lower bound of $k + \Omega(\sqrt{k})$.

Similarly, standard online learning lower bound techniques also yield a lower bound for DFF learning with deterministic DFF learners.

Observation 8 *Consider a class with DFF-dimension $d \geq 1$. Every deterministic learner has a worst-case expected mistake bound of at least $2k + d$.*

In a class with DFF-dimension $d \geq 1$, the class has at least two teachers whose labeling differs on at least one point x. In each round t, the adversary can present the point $x_t = x$. since the algorithm is deterministic, the adversary can select a teacher (f_t, ψ_t) with $f_t(x_t) = y_t \neq \hat{y}_t$. Thus, the algorithm will make a mistake in round t. This can continue in all rounds, as long as one of the two teachers is consistent with all but at most k rounds. This is the case at the earliest after $2k + 1$ rounds. After this, the worst case algorithm can force $d - 1$ additional mistakes, due to the DFF-dimension being d.

In the next sections, we provide bounds for specific classes that prove separation results.

5.1. A mistake upper bound for $\mathcal{T}_3$ by a randomized learner with an oblivious adversary

In this section, we will show that for oblivious adversaries, there exists a class that can be learned with vanishing regret by randomized learners, while deterministic learners can be forced by an adversary to make a mistake in every round.

Theorem 9 *For the class $\mathcal{T}_3$ and history H_3 the following two statements are true:*

- *$(\mathcal{T}_3, H_3)$ has no finite mistake bound in the deterministic realizable case.*
- *$(\mathcal{T}_3, H_3)$ can be learned with vanishing regret by a randomized algorithm in the case with oblivious adversaries. In particular, for a run with T rounds, the expected mistake bound for oblivious adversaries is at most $k + O\left(\sqrt{k \log k} + \log T\right)$.*

The first part of this result follows from Lemma 5 in Section 4. The second part is proved in Lemma 10 below, using Algorithm 1. This algorithm maintains experts with weights and multiplicative updates similarly to the Weighted Majority (WM) algorithm (Littlestone and Warmuth, 1994). In contrast to the standard WM algorithm, it maintains a growing set of experts E_t, consisting only of teachers that have been revealed to the learner through feature feedback, rather than a the whole set of possible teachers. In addition, in order to make sure that the true teacher is revealed through feature feedback, the algorithm keeps predicting a label uniformly at random in "exploration rounds". In those rounds, with probability $1/3$ a new teacher is revealed and added to the set of experts. In non-exploration rounds, the algorithm selects a random expert according to their weights and uses it to predict.

The probability of an exploration round in round t is given by a parameter γ_t. We assume that there is a cut-off T_0, such that $\gamma_t = 0$ for $t \geq T_0$. The multiplicative updates are controlled by a parameter β. We can show that this algorithm obtains the following bound.

Lemma 10 *Fix $T_0 \geq 1$ and run Algorithm 1 with $\gamma_t = \frac{1}{\sqrt{t}} \mathbf{1}\{t \leq T_0\}$ and $\beta = 1 - \sqrt{\frac{\log(k+1)}{k}}$, for some $k \geq 10$ (so that $\beta \geq 1/2$). Let M_T be the number of mistakes until round T. Then for every horizon $T \geq 1$,*

$$\mathbb{E}[M_T] \leq k + 2\sqrt{k \log(k+1)} + 2\sqrt{T_0} + 3\sqrt{k+1} + \frac{11}{2} + T \exp\left(-\frac{2}{3}\left(\sqrt{T_0} - \sqrt{k+1}\right)\right).$$

We now give a proof sketch of this result. The full proof can be found in Appendix A. For all $t \geq 1$, we define

$$\varepsilon_t := \sum_{h \in E_t : h(x_t) \neq y_t} \frac{w_t(h)}{Z_t}.$$

Algorithm 1 Exploration + multiplicative weights with a growing expert set and cutoff T_0

1: Input: $T, k, \{\gamma_t\}_{t\geq 1}$
2: Initialize: $E_1 = \{h_1\}$ and $w_1(h_1) = 1$.
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Compute $Z_t = \sum_{h\in E_t} w_t(h)$.
5: Sample $b_t \sim \text{Bernoulli}(\gamma_t)$.
6: **if** $b_t = 1$ **then**
7: Predict $\hat{y}_t$ uniformly from $\mathcal{Y}$.
8: **else**
9: Sample $h \in E_t$ with probability $w_t(h)/Z_t$.
10: Predict $\hat{y}_t = h(x_t)$.
11: Observe y_t. If $\hat{y}_t \neq y_t$, obtain feature feedback ϕ_t
12: Update weights for all $h \in E_t$: $w_{t+1}(h) \leftarrow w_t(h)\beta^{\mathbf{1}\{h(x_t)\neq y_t\}}$.
13: Let $Z^{\text{old}}_{t+1} \leftarrow \sum_{h\in E_t} w_{t+1}(h)$.
14: **if** ϕ_t reveals new expert **then**
15: Set $E_{t+1} \leftarrow E_t \cup \{h'\}$.
16: Set $w_{t+1}(h') \leftarrow Z^{\text{old}}_{t+1}/|E_t|$.
17: **else**
18: Set $E_{t+1} \leftarrow E_t$.

When a new expert is added, it is assigned the average weight of all current experts (line 16 in the algorithm). Due to this fact, we can obtain the following formula for Z_{t+1}, which follows from a similar analysis as the one for standard multiplicative weight algorithm with a prefixed set of experts.

Lemma 11 *For every $t \geq 1$, we have $Z_{t+1} = |E_{t+1}| \prod_{i=1}^{t}(1 - (1-\beta)\varepsilon_i)$.*

Next, we observe that since the adversary is oblivious, we may assume that there is an unknown true teacher $h^* = (f^*, \psi_{f^*})$ and the adversary fixes in advance a set of rounds which are exception rounds. We define τ as the random variable that indicates the earliest round in which the true teacher is included in the expert class. This means that the feature feedback in round $\tau - 1$ is $\phi_{\tau-1} = \phi_{f^*,x_{\tau-1},\hat{x}_{\tau-1-1}}$. Thus h^* is in E_τ. Using this, we can obtain a mistake bound under the assumption that the true expert has been revealed before the cut-off T_0.

Lemma 12 *Assume $\tau \leq T_0$. Then*

$$\sum_{t=\tau}^{T} \varepsilon_t \leq \frac{\log(1/\beta)}{1-\beta}(k - k_{\tau-1}) + \frac{\log(k+1)}{1-\beta}.$$

Lastly, we bound the probability that the the true teacher is revealed before round T_0.

Lemma 13 *Fix $c \in (0, 1]$ and set $\gamma_t = \frac{c}{\sqrt{t}}\mathbf{1}\{t \leq T_0\}$. Then*

$$\mathbb{P}(\tau > T_0) \leq \exp\left(-\frac{2c}{3}(\sqrt{T_0} - \sqrt{k+1})\right).$$

With Lemma 10 we can now prove Theorem 9, by setting $T_0 = \left\lceil \sqrt{k+1} + \frac{3}{2}\log T \right\rceil^2 - 1$. Full proofs of Lemma 11, Lemma 12 and Lemma 13, Theorem 9, can be found in Appendix A.

5.2. Separation results for adaptive adversaries

In this section, we focus on learning with randomized algorithms in the presence of an adaptive adversary. We show two separation results. First, we show that there are classes that are not learnable by deterministic learners (not even in the realizable case), but can be have a non-trivial mistake bound in the non-realizable case by randomized algorithms against an adaptive adversary, although we do not obtain vanishing regret.

Theorem 14 *Consider the class $\mathcal{T}_{[0,1]}$ with the infinite history $H_{[0,1]}$, defined in Section 4. No deterministic learner has a finite mistake bound for this pair (not even in the realizable setting), but the class can be learned by a randomized algorithm against an adaptive adversary with a mistake bound satisfying*

$$\mathcal{M}_k^{\text{adapt}}(\mathcal{A}, \mathcal{T}_{[0,1]}, H_{[0,1]}) \leq 2k + O(\sqrt{k(\log(k))}).$$

Proof Lemma 3 shows that the class $\mathcal{T}_{[0,1]}$ and history $H_{[0,1]}$ cannot be learned with a finite mistake bound by any deterministic learner, not even in the realizable setting. It remains to show that this class can be learned by a randomized learner in the presence of an adaptive adversary with the mistake bound above.

Consider the following randomized learner: if $x_t \notin [0,1]$ in any round t then the learner can predict $g(x_t)$ which is the correct label. Therefore, assume without loss of generality that $x_t \in [0,1]$ in all rounds. In every round $t \leq k+1$, the learner selects a label $\hat{y}_t$ uniformly at random from the set $[0,1]$. It then predicts $(\hat{y}_t, -\hat{y}_t)$. If in such a round t, the adversary has selected a teacher from the class, $(f_t, \psi_{f_t}) \in \mathcal{T}_{[0,1]}$, then $y_t = f_t(x_t)$ and with probability 1, $y_t \neq \hat{y}_t \neq (y_t + \frac{1}{2}) \bmod 1$, which means a prediction mistake will occur and ϕ_{f_t,x_t} will be provided as feedback. We define $\mathcal{F}' = \{f \in \mathcal{F} \mid \exists t \leq k+1 \text{ s.t. } \phi_t = \phi_{f,x_t}\}$. Clearly, $|\mathcal{F}'| = k+1$, and since there can only be k exceptions in the sequence, one of the teachers (f_t, ψ_{f_t}) with $f_t \in \mathcal{F}'$ corresponds to a teacher which is consistent with all but at most k rounds. After the first $k+1$ rounds, the algorithm runs the Weighted Majority algorithm for Online Learning (Littlestone and Warmuth, 1994) on the class $\mathcal{F}'$, giving a mistake bound of $k + O(\sqrt{k \log k})$ on that part of the run. Taking together the mistakes made until round $k+1$ and the mistakes after round $k+1$, we get a mistake bound of

$$\mathcal{M}_k^{\text{adapt}}(\mathcal{A}, \mathcal{T}_{[0,1]}, H_{[0,1]}) \leq (k+1) + k + O(k \log(k+1)) = 2k + O(\sqrt{k \log k}).$$

■

Our second separation result shows that the class $\mathcal{T}_3$ studied above cannot have a vanishing regret with an adaptive adversary, in contrast to the oblivious adversary, which was studied in Theorem 9 above. This is unlike Online Learning, where the oblivious and adaptive adversaries obtain essentially the same optimal mistake bounds for all classes.

Theorem 15 *Consider $\mathcal{T}_3$ with history H_3. Let $T > 6k + 8$ and $k \geq 2$. For any algorithm $\mathcal{A}$, we have $\mathcal{M}_k^{\text{adapt}}(\mathcal{T}_3, H_3) \geq 2k-2$. In contrast, there exists an algorithm $\mathcal{A}$ such that $\mathcal{M}_k^{\text{obl}}(\mathcal{A}, \mathcal{T}_3, H_3) = k + O\left(\sqrt{k \log k} + \log T\right)$.*

The lower bound for the adaptive adversary follows from the fact that an adaptive adversary is allowed to decide in retrospect on the "true teacher". Therefore, it is allowed to only commit to a teacher after $k+1$ teachers are revealed and render the first k rounds in which teachers were revealed "exceptions". In expectation, the algorithm will make roughly $2k$ mistakes until such a time, since only roughly one half of the mistakes that are made reveal a teacher. The full proof can be found in Appendix B.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) [funding reference number RGPIN-2024-05907]. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute; see https://vectorinstitute.ai/partnerships/current-partners/. This work was done while Valentio Iverson was a Research Intern at the Vector Institute.

References

Raman Arora, Ofer Dekel, and Ambuj Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1747–1754, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Omri Bar Oz, Tosca Lechner, and Sivan Sabato. Discriminative feature feedback with general teacher classes. In *37th International Conference on Algorithmic Learning Theory*, 2025.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon Van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 532–541, 2021.

W.B. Croft and R. Das. Experiments with query acquisition and use in document retrieval systems. In *Proceedings of the 13th International Conference on Research and Development in Information Retrieval*, pages 349–368, 1990.

Sanjoy Dasgupta and Sivan Sabato. Robust learning from discriminative feature feedback. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 973–982. PMLR, 2020.

Sanjoy Dasgupta, Akansha Dey, Nicholas Roberts, and Sivan Sabato. Learning from discriminative feature feedback. In *Advances in Neural Information Processing Systems 31*, pages 3955–3963. Curran Associates, Inc., 2018.

G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of ACM Special Interest Group on Information Retrieval*, 2008.

Yuval Filmus, Steve Hanneke, Idan Mehalel, and Shay Moran. Optimal prediction using expert advice and randomized littlestone dimension. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 773–836. PMLR, 2023.

Steve Hanneke, Amin Karbasi, Shay Moran, and Grigoris Velegkas. Universal rates for interactive learning. *Advances in Neural Information Processing Systems*, 35:28657–28669, 2022.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563. Association for Computational Linguistics, December 2022.

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2:285–318, 1988.

N. Littlestone and M.K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. ISSN 0890-5401.

O. Mac Aodha, S. Su, Y. Chen, P. Perona, and Y. Yue. Teaching categories to human learners with visual explanations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, 56(4):3005–3054, August 2022. ISSN 0269-2821.

H. Raghavan, O. Madani, and R. Jones. Interactive feature selection. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 841–846, 2005.

Sivan Sabato. Improved robust algorithms for learning with discriminative feature feedback. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 1024–1036. PMLR, 2023.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

B. Settles. Closing the loop: fast, interactive semi-supervised annotation with queries on features and instances. In *Empirical Methods in Natural Language Processing*, 2011.

Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 239–245, 2019.

Chhavi Yadav, Michal Moshkovitz, and Kamalika Chaudhuri. XAudit : A learning-theoretic look at auditing with explanations. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Appendix A. Proof of Theorem 9

We observe that since the adversary is oblivious, we may assume that there is an unknown true expert h^*, and the adversary fixes in advance a set $B \subseteq \{1, 2, 3, \dots\}$ with $|B| \leq k$, which will be the *exception rounds*, in which it responds in a way that may not be consistent with h^*. Note that by the definition of $\mathcal{T}_3$, if $t \notin B$, there exists at least one label $a_t \in \mathcal{Y}$ such that if the learner predicts

$\hat{y}_t = a_t$ at round t, then at the end of round t, the adversary reveals the correct expert h^*. Given $(E_t)_{t\geq 1}$, where E_t is the set of available experts at the start of round t, We define

$$\tau := \inf\{t \geq 1 : h^* \in E_t\} \in \{1, 2, \dots\} \cup \{\infty\}$$

In this setup, the environment may reveal new experts to the learner.

As there are at most k exception rounds, there could be at most k additional distinct experts being introduced, as every wrong expert introduced happens in an exception round. In particular, for every time t, we have $|E_t| \leq k + 1$. For any time $t = 1, 2, \dots, T$, we define

$$G_t := \{s \in \{1, 2, \dots, t\} : s \notin B\}$$
$$B_t := \{s \in \{1, 2, \dots, t\} : s \in B\}$$

to be the set of non-exception and exception rounds up till time t, and we denote $k_t := |B_t|$. Note that by definition $k_\tau \leq k$ holds.

Remember that for $t \geq 1$, we defined

$$\varepsilon_t = \sum_{h \in E_t : h(x_t) \neq y_t} \frac{w_t(h)}{Z_t}.$$

We will now prove Lemma 11, which stated that for every $t \geq 1$, we have

$$Z_{t+1} = |E_{t+1}| \prod_{i=1}^{t} (1 - (1 - \beta)\varepsilon_i).$$

Proof [of Lemma 11] We can prove this by induction on t. For $t = 1$, this is clearly true. Now, we see that by the update at round t, we have

$$\begin{aligned} Z_{t+1}^{\text{old}} &= \sum_{h \in E_t} w_t(h) \beta^{\mathbf{1}\{h(x_t) \neq y_t\}} \\ &= \sum_{h \in E_t :\, h(x_t) = y_t} w_t(h) + \beta \sum_{h \in E_t :\, h(x_t) \neq y_t} w_t(h) \\ &= Z_t \left((1 - \varepsilon_t) + \beta \varepsilon_t\right) \\ &= Z_t \left(1 - (1 - \beta)\varepsilon_t\right). \end{aligned}$$

Now, there are two cases: if there are no new experts added after round t, we have $Z_{t+1} = Z_{t+1}^{\text{old}}$ and $|E_{t+1}| = |E_t|$ and thus

$$Z_{t+1} = Z_t(1 - (1 - \beta)\varepsilon_t).$$

If a new expert is added after round t, note that each round only at most one new expert can be introduced: thus $|E_{t+1}| = |E_t| + 1$ in this case and thus by our initialization rule,

$$Z_{t+1} = Z_{t+1}^{\text{old}} + \frac{1}{|E_t|} Z_{t+1}^{\text{old}} = \frac{|E_{t+1}|}{|E_t|} Z_t \left(1 - (1 - \beta)\varepsilon_t\right) = |E_{t+1}| \prod_{i=1}^{t} (1 - (1 - \beta)\varepsilon_i).$$

We note that here we use the inductive hypothesis on time t, which says that

$$Z_t = |E_t| \prod_{i=1}^{t-1} (1 - (1 - \beta)\varepsilon_i).$$

■

We will now compare the weighted error after τ to the performance of h^* and prove Lemma 12, which stated that for $\tau \leq T_0$, we have

$$\sum_{t=\tau}^{T} \varepsilon_t \leq \frac{\log(1/\beta)}{1-\beta}(k - k_{\tau-1}) + \frac{\log(k+1)}{1-\beta}$$

Proof [of Lemma 12] By definition of τ, we know that $h^* \in E_\tau$ and $h^* \notin E_{\tau-1}$. At the end of round $\tau - 1$, the algorithm assigns the initial weight

$$w_\tau(h^*) = \frac{1}{|E_{\tau-1}|} Z_\tau^{\text{old}}.$$

We substitute the update rule $Z_\tau^{\text{old}} = Z_{\tau-1}(1 - (1-\beta)\varepsilon_{\tau-1})$. Furthermore, applying Lemma 2.1 at time $\tau - 1$ gives us $Z_{\tau-1} = |E_{\tau-1}| \prod_{i=1}^{\tau-2}(1 - (1-\beta)\varepsilon_i)$. Combining these, we observe that the $|E_{\tau-1}|$ terms cancel out:

$$w_\tau(h^*) = \frac{1}{|E_{\tau-1}|} \cdot |E_{\tau-1}| \prod_{i=1}^{\tau-2}(1 - (1-\beta)\varepsilon_i) \cdot (1 - (1-\beta)\varepsilon_{\tau-1}) = \prod_{i=1}^{\tau-1}(1 - (1-\beta)\varepsilon_i).$$

From time τ onwards, h^* is penalized only when the round is an exception round. The number of exception rounds in the interval $[\tau, T]$ is exactly $k_T - k_{\tau-1}$. Since $k_T \leq k$, there are at most $k - k_{\tau-1}$ such rounds. Thus, the final weight satisfies:

$$w_{T+1}(h^*) \geq w_\tau(h^*)\beta^{k-k_{\tau-1}} = \beta^{k-k_{\tau-1}} \prod_{i=1}^{\tau-1}(1 - (1-\beta)\varepsilon_i).$$

Now Lemma 2.1 gives us the total mass at the end:

$$Z_{T+1} = |E_{T+1}| \prod_{i=1}^{T}(1 - (1-\beta)\varepsilon_i).$$

Since the total mass must be at least the weight of one expert, we have $Z_{T+1} \geq w_{T+1}(h^*)$. Now note that we must have $|E_{T+1}| \leq k + 1$ as a new expert, other than the true expert can only be introduced in an exception round, which happens at most k times. This yields:

$$\beta^{k-k_{\tau-1}} \prod_{i=1}^{\tau-1}(1 - (1-\beta)\varepsilon_i) \leq (k+1) \prod_{i=1}^{T}(1 - (1-\beta)\varepsilon_i).$$

Simplifying this, gives us

$$\beta^{k-k_{\tau-1}} \leq (k+1) \prod_{t=\tau}^{T}(1 - (1-\beta)\varepsilon_t).$$

Taking logarithms on both sides and using the inequality $\ln(1-x) \leq -x$, we get

$$(k - k_{\tau-1}) \log \beta \leq \log(k+1) - (1-\beta) \sum_{t=\tau}^{T} \varepsilon_t.$$

Rearranging the terms and dividing by $(1-\beta)$ gives the result immediately. ■

To streamline our analysis, we now establish a general concentration inequality regarding the number of non-exception rounds required to reveal the true expert. Recall that $G_{\tau-1}$ denotes the set of non-exception rounds strictly prior to the discovery of the true expert h^*.

Lemma 16 *Fix $c \in (0,1]$ and set $\gamma_t = \frac{c}{\sqrt{t}}\mathbf{1}\{t \le T_0\}$. For any integer $m \ge 1$, if the m-th non-exception round occurs by time T_0, then*

$$\mathbb{P}(|G_{\tau-1}| \ge m) \le \exp\left(-\frac{2c}{3}(\sqrt{k+m} - \sqrt{k+1})\right).$$

Proof Let $t_1 < t_2 < t_3 < \cdots$ be the increasing sequence of non-exception rounds, i.e., rounds t such that $t \notin B$. Since $|B| \le k$, among the first $k+i$ rounds there are at most k exception rounds, hence at least i non-exception rounds. Therefore, $t_i \le k+i$ for all $i \ge 1$.

Let $(\mathcal{H}_t)_{t\ge 0}$ be the history up to round t, including learner's internal randomness, labels and experts. Fix an index $i \ge 1$. On the event $\{\tau > t_i\}$, we have $h^* \notin E_{t_i}$ at the start of round t_i. Since $t_i \notin B$, by the problem setup there exists a label $a_{t_i} \in \mathcal{Y}$ such that if the learner predicts $\hat{y}_{t_i} = a_{t_i}$, then the environment reveals h^* at the end of round t_i, which implies $\tau \le t_i + 1$.

At round t_i, the learner explores with probability γ_{t_i} and, conditional on exploring, predicts uniformly from $\mathcal{Y}$, hence

$$\mathbb{P}(\hat{y}_{t_i} = a_{t_i} \mid \mathcal{H}_{t_i-1}) \ge \frac{\gamma_{t_i}}{3}.$$

and hence we obtain the conditional probability

$$\mathbb{P}(\tau > t_i + 1 \mid \tau > t_i) \le 1 - \frac{\gamma_{t_i}}{3}.$$

The event $|G_{\tau-1}| \ge m$ means that the discovery round $\tau - 1$ occurs at or after the m-th non-exception round. Equivalently,

$$\{|G_{\tau-1}| \ge m\} \subseteq \{\tau > t_{m-1} + 1\},$$

where we use the convention $t_0 := 0$. By repeated application of the above one-step bound,

$$\mathbb{P}(|G_{\tau-1}| \ge m) \le \mathbb{P}(\tau > t_{m-1}+1) \le \prod_{i=1}^{m-1}\left(1 - \frac{\gamma_{t_i}}{3}\right) \le \exp\left(-\frac{1}{3}\sum_{i=1}^{m-1}\gamma_{t_i}\right),$$

where we used $1 - x \le e^{-x}$.

Assume now that the m-th non-exception round occurs by time T_0, so $t_m \le T_0$. Then $\gamma_{t_i} = c/\sqrt{t_i}$ for all $1 \le i \le m-1$. Since $t \mapsto c/\sqrt{t}$ is decreasing and $t_i \le k+i$, we have

$$\gamma_{t_i} \ge \frac{c}{\sqrt{k+i}}.$$

Therefore,

$$\sum_{i=1}^{m-1}\gamma_{t_i} \ge c\sum_{i=1}^{m-1}\frac{1}{\sqrt{k+i}}.$$

Using the integral comparison,

$$\sum_{i=1}^{m-1} \frac{1}{\sqrt{k+i}} \geq \int_{k+1}^{k+m} \frac{1}{\sqrt{x}}\, dx = 2(\sqrt{k+m} - \sqrt{k+1}).$$

Substituting into the exponential bound gives

$$\mathbb{P}(|G_{\tau-1}| \geq m) \leq \exp\left(-\frac{2c}{3}(\sqrt{k+m} - \sqrt{k+1})\right),$$

as desired. ■

We will now prove Lemma 13, which states that for fixed $c \in (0,1]$ and $\gamma_t = \frac{c}{\sqrt{t}}\mathbf{1}\{t \leq T_0\}$, we get the following bound on the probability for $\tau > T_0$:

$$\mathbb{P}(\tau > T_0) \leq \exp\left(-\frac{2c}{3}(\sqrt{T_0} - \sqrt{k+1})\right).$$

Proof [of Lemma 13] Let $m_0 = |G_{T_0-1}|$, the number of non-exception rounds up to time $T_0 - 1$. If $\tau > T_0$, then $h^* \notin E_{T_0}$, which means that h^* was not revealed by the end of round $T_0 - 1$. Hence the discovery round $\tau - 1$ occurs after all non-exception rounds up to time $T_0 - 1$, and therefore

$$\{\tau > T_0\} \subseteq \{|G_{\tau-1}| \geq m_0 + 1\}.$$

Therefore,

$$\mathbb{P}(\tau > T_0) \leq \mathbb{P}(|G_{\tau-1}| \geq m_0 + 1).$$

Moreover, among the first $T_0 - 1$ rounds there are at most $|B| \leq k$ exception rounds, so

$$T_0 - 1 \leq |B| + m_0 \leq k + m_0,$$

and thus $T_0 \leq k + (m_0 + 1)$, which implies

$$\sqrt{k + m_0 + 1} \geq \sqrt{T_0}.$$

If $m_0 + 1 \geq 1$, then the $(m_0 + 1)$-th non-exception round occurs by time T_0, so Lemma 16 applies with $m = m_0 + 1$ and yields

$$\mathbb{P}(|G_{\tau-1}| \geq m_0+1) \leq \exp\left(-\frac{2c}{3}(\sqrt{k + m_0 + 1} - \sqrt{k+1})\right) \leq \exp\left(-\frac{2c}{3}(\sqrt{T_0} - \sqrt{k+1})\right).$$

■

We will now show a precursor of Lemma 10.

Lemma 17 *Fix $c \in (0,1]$ and set $\gamma_t = \frac{c}{\sqrt{t}}\mathbf{1}\{t \leq T_0\}$. Then*

$$\mathbb{E}\big[|G_{\tau-1}|\,\mathbf{1}\{\tau \leq T_0\}\big] \leq \frac{3}{c}\sqrt{k+1} + \frac{9}{2c^2}.$$

Proof Let $X := |G_{\tau-1}|$ and $A := \{\tau \le T_0\}$. Since X is a nonnegative integer-valued random variable, the tail-sum formula gives

$$\mathbb{E}[X\mathbf{1}\{A\}] = \sum_{m\ge 1} \mathbb{P}(X \ge m, A).$$

Fix $m \ge 1$ and let t_m be the time of the m-th non-exception round (deterministic given B). If $t_m > T_0$, then the event $\{X \ge m, A\}$ is empty, because $X \ge m$ implies $\tau \ge t_m + 1 > T_0$, contradicting A. Hence $\mathbb{P}(X \ge m, A) = 0$ in this case.

If instead $t_m \le T_0$, then Lemma 16 applies and yields

$$\mathbb{P}(X \ge m, A) \le \mathbb{P}(X \ge m) \le \exp\left(-\frac{2c}{3}(\sqrt{k+m} - \sqrt{k+1})\right).$$

Summing over m gives

$$\mathbb{E}[X\mathbf{1}\{A\}] \le \sum_{m=1}^{\infty} \exp\left(-\frac{2c}{3}(\sqrt{k+m} - \sqrt{k+1})\right).$$

We now bound the sum as follow by an integral:

$$\mathbb{E}[X\mathbf{1}\{A\}] \le \int_0^{\infty} \exp\left(-\frac{2c}{3}(\sqrt{k+u} - \sqrt{k+1})\right) du.$$

With the change of variables $v = \sqrt{k+u} - \sqrt{k+1}$, we have $u = v^2 + 2v\sqrt{k+1}$ and $du = (2v + 2\sqrt{k+1})\,dv$, so the integral equals

$$\begin{aligned}\int_0^{\infty} e^{-\frac{2c}{3}v}(2v + 2\sqrt{k+1})\,dv &= 2\int_0^{\infty} ve^{-\frac{2c}{3}v}\,dv + 2\sqrt{k+1}\int_0^{\infty} e^{-\frac{2c}{3}v}\,dv \\ &= 2\cdot\frac{1}{(2c/3)^2} + 2\sqrt{k+1}\cdot\frac{1}{2c/3} \\ &= \frac{9}{2c^2} + \frac{3\sqrt{k+1}}{c}.\end{aligned}$$

This completes the proof. ■

Now, we are well equipped to prove the main lemma, Lemma 10. Recall that this lemma gives a mistake bound of

$$\mathbb{E}[M_T] \le k + 2\sqrt{k\log(k+1)} + 2\sqrt{T_0} + 3\sqrt{k+1} + \frac{11}{2} + T\exp\left(-\frac{2}{3}(\sqrt{T_0} - \sqrt{k+1})\right),$$

Proof [of Lemma 10] Let $I_t := \mathbf{1}\{\hat{y}_t \ne y_t\}$, so that $M_T = \sum_{t=1}^{T} I_t$. Let $\mathcal{H}_{t-1}$ be the history up to the end of round $t-1$, capturing learner's internal randomness, labels and revealed experts. Recall that we defined

$$\varepsilon_t := \sum_{h\in E_t : h(x_t)\ne y_t} \frac{w_t(h)}{Z_t}.$$

Condition on $\mathcal{H}_{t-1}$ and on (x_t, y_t). With probability γ_t the learner explores and makes a mistake with probability at most 1. With probability $1-\gamma_t$ the learner samples $h \in E_t$ with probability

$w_t(h)/Z_t$ and predicts $h(x_t)$, in which case the conditional mistake probability is exactly ε_t. Therefore, $\mathbb{E}[I_t \mid \mathcal{H}_{t-1}, x_t, y_t] \le \gamma_t + \varepsilon_t$. Taking expectations and summing over t yields

$$\mathbb{E}[M_T] \le \sum_{t=1}^{T} \gamma_t + \mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t\right]. \tag{1}$$

Since $\gamma_t = 0$ for $t > T_0$ and $\sum_{t=1}^{T_0} 1/\sqrt{t} \le 2\sqrt{T_0}$, we have

$$\sum_{t=1}^{T} \gamma_t \le 2\sqrt{T_0}.$$

Let $\mathcal{E} := \{\tau \le T_0\}$. Then

$$\mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t\right] = \mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t \, \mathbf{1}\{\mathcal{E}\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t \, \mathbf{1}\{\mathcal{E}^c\}\right].$$

On $\mathcal{E}^c$ we use the trivial bound $\varepsilon_t \le 1$ for each t, hence

$$\mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t \, \mathbf{1}\{\mathcal{E}^c\}\right] \le T\mathbb{P}(\mathcal{E}^c).$$

By Lemma 13 with $c = 1$,

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}(\tau > T_0) \le \exp\left(-\frac{2}{3}\left(\sqrt{T_0} - \sqrt{k+1}\right)\right),$$

so

$$\mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t \, \mathbf{1}\{\mathcal{E}^c\}\right] \le T \exp\left(-\frac{2}{3}\left(\sqrt{T_0} - \sqrt{k+1}\right)\right).$$

On $\mathcal{E}$ we have $\tau \le T_0$, and we split at τ:

$$\sum_{t=1}^{T} \varepsilon_t = \sum_{t=1}^{\tau-1} \varepsilon_t + \sum_{t=\tau}^{T} \varepsilon_t.$$

For the pre-discovery part, we use $\varepsilon_t \le 1$ and the identity $\tau - 1 = |G_{\tau-1}| + |B_{\tau-1}|$:

$$\sum_{t=1}^{\tau-1} \varepsilon_t \le \tau - 1 = |G_{\tau-1}| + k_{\tau-1}.$$

For the post-discovery part, Lemma 2.2 yields on $\mathcal{E}$:

$$\sum_{t=\tau}^{T} \varepsilon_t \le \frac{\log(1/\beta)}{1-\beta}(k - k_{\tau-1}) + \frac{\log(k+1)}{1-\beta}.$$

Let $\alpha := 1 - \beta \in (0, 1/2]$. Using $\log(1/(1-\alpha)) \le \alpha + \alpha^2$ for $\alpha \in (0, 1/2]$ gives

$$\frac{\log(1/\beta)}{1-\beta} = \frac{\log(1/(1-\alpha))}{\alpha} \le 1 + \alpha = 2 - \beta.$$

Thus, we obtain

$$\sum_{t=\tau}^{T} \varepsilon_t \le (2-\beta)(k-k_{\tau-1}) + \frac{\log(k+1)}{1-\beta}$$

Combining the two parts, on $\mathcal{E}$ we obtain

$$\begin{aligned}
\sum_{t=1}^{T} \varepsilon_t &\le |G_{\tau-1}| + k_{\tau-1} + (2-\beta)(k-k_{\tau-1}) + \frac{\log(k+1)}{1-\beta} \\
&\le |G_{\tau-1}| + k + (1-\beta)(k-k_{\tau-1}) + \frac{\log(k+1)}{1-\beta} \\
&\le |G_{\tau-1}| + k + (1-\beta)k + \frac{\log(k+1)}{1-\beta} \\
&\le |G_{\tau-1}| + k + 2\sqrt{k\log(k+1)}
\end{aligned}$$

with our choice of $\beta = 1 - \sqrt{\log(k+1)/k}$. Multiplying by $\mathbf{1}\{\mathcal{E}\}$ and taking expectations gives

$$\mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t \mathbf{1}\{\mathcal{E}\}\right] \le k + 2\sqrt{k\log(k+1)} + \mathbb{E}\big[|G_{\tau-1}|\mathbf{1}\{\mathcal{E}\}\big].$$

Applying Lemma 17 with $c = 1$ yields

$$\mathbb{E}\big[|G_{\tau-1}|\mathbf{1}\{\mathcal{E}\}\big] \le 1 + 3\sqrt{k+1} + \frac{9}{2}.$$

Therefore,

$$\mathbb{E}\left[\sum_{t=1}^{T} \varepsilon_t\right] \le k + 2\sqrt{k\log(k+1)} + 3\sqrt{k+1} + \frac{11}{2} + T\exp\left(-\frac{2}{3}\big(\sqrt{T_0} - \sqrt{k+1}\big)\right).$$

Substituting the above estimate and $\sum_{t=1}^{T} \gamma_t \le 2\sqrt{T_0}$ into (1) completes the proof. ■

From this we get a corollary which states the upper bound for oblivious learners in Theorem 9.

Corollary 18 *Assume $k \ge 10$. Fix a horizon $T \ge 1$ and run Algorithm 1 with*

$$\gamma_t = \frac{1}{\sqrt{t}}\mathbf{1}\{t \le T_0\} \quad \textit{and} \quad \beta = 1 - \sqrt{\frac{\log(k+1)}{k}},$$

where

$$T_0 = \left\lceil \sqrt{k+1} + \frac{3}{2}\log T \right\rceil^2 - 1.$$

Then

$$\mathbb{E}[M_T] \le k + 2\sqrt{k\log(k+1)} + 5\sqrt{k+1} + 3\log T + \frac{19}{2} = k + O\left(\sqrt{k\log k} + \log T\right).$$

Proof By the definition of T_0,

$$\sqrt{T_0+1} = \left\lceil \sqrt{k+1} + \frac{3}{2}\log T \right\rceil \geq \sqrt{k+1} + \frac{3}{2}\log T,$$

and therefore

$$\sqrt{T_0} \geq \sqrt{k+1} + \frac{3}{2}\log T - 1,$$

which implies

$$T\exp\left(-\frac{2}{3}\left(\sqrt{T_0}-\sqrt{k+1}\right)\right) \leq T\exp\left(-\frac{2}{3}\left(\frac{3}{2}\log T - 1\right)\right) = e^{2/3}.$$

Substituting these two bounds into the conclusion of the main theorem yields the stated inequality. ■

We can now prove the main theorem, which stated that the class $\mathcal{T}_3$ with history H_3 satisfies the following two properties:

- $(\mathcal{T}_3, H_3)$ has no finite mistake bound in the deterministic realizable case.
- $(\mathcal{T}_3, H_3)$ can be learned with vanishing regret by a randomized algorithm in the case with oblivious adversaries. In particular, for a run wit T rounds, the expected mistake bound for oblivious adversaries is at most $k + O\left(\sqrt{k\log k} + \log T\right)$.

Proof [of Theorem 9] The first part of the Theorem follows from Lemma 5. The second part of the theorem follows from Corollary 18. ■

Appendix B. Proof of Theorem 15

Proof We show that there exists an adaptive adversary that causes every learner to make at least $2k-2$ mistakes in expectation. In a given run, denote the mistake rounds with potentially "useful feature feedback", by $C_t = \{s \leq t \mid y_s - \hat{y}_s \equiv 2 \pmod 3\}$.

In every round t, if $|C_{t-1}| \leq k+1$, the adversary selects $x_t = t+3$ and $y_t = \text{uniform}(0,1,2)$. Furthermore, it selects a teacher $h_t = (f, \psi_f) \in \mathcal{T}_3$ such that f is consistent with $H_3 \cup \{(x_i, y_i)\}_{i\leq t, i\notin C_t}$.

If $|C_{t-1}| = k+1$ and this is the smallest such t, consider the teacher h_{t-1} that was selected in round $t-1$. Define $h^* = (f^*, \psi_{f^*}) := h_{t-1}$. Furthermore, define the set $E := C_{t-1}$ and let $\tau := t-1$. In every round with $|C_{t-1}| > k+1$, the adversary selects $x_t = t+3$, $y_t = f^*(x_t)$ and teacher $h_t = h^*$.

We note that every round not in $t \in [T] \setminus E$ as well as one round in E, the round is consistent with the teacher h^*. Thus, there are at most $|E| - 1 = k$ exceptions for this adversary, consistent with a k-non-realizable setting. We will now argue that every algorithm makes at least $2k-2$ mistakes in expectation.

For every round t, conditioned on $t \leq \tau$, for every algorithm, with probability $1/3$ the round is a mistake round with "potentially useful" feedback, that is $y_t - \hat{y}_t \equiv 2 \pmod 3$, and with probability $1/3$ the round is a mistake round with an "un-useful" feedback, that is $y_t - \hat{y}_t \equiv 1 \pmod 3$.

Let $R_t = 1_{\{y_t - \hat{y}_t \equiv 1 \pmod 3\}}$ indicate whether round t has an un-useful mistake. If τ is defined in the run, the total number of mistakes is at least $k + 1 + \sum_{t=1}^{\tau} R_t$. Therefore,

$$\mathcal{M}_k^{\text{adapt}}(\mathcal{T}_3, H_3) \geq \mathbb{E}[\#\text{mistakes until round } \tau \mid \tau \text{ is defined}] \cdot \mathbb{P}(\tau \text{ is defined}) \tag{2}$$

$$= (k + 1 + \mathbb{E}[\sum_{t=1}^{\tau} R_t]) \cdot \mathbb{P}(\tau \text{ is defined}). \tag{3}$$

We now bound the expected sum:

$$\begin{aligned}
\mathbb{E}[\sum_{t=1}^{T} R_t] &= \mathbb{E}[\sum_{t=1}^{T} R_t \cdot 1_{\{t<\tau\}}] = \sum_{t=1}^{T} \mathbb{P}(R_t \mid t < \tau) \cdot \mathbb{P}(t < \tau) \\
&= \sum_{t=1}^{T} \tfrac{1}{3}\mathbb{P}(t < \tau) = \tfrac{1}{3} \sum_{t=1}^{T} \mathbb{P}(t < \tau) \\
&= \tfrac{1}{3}(\mathbb{E}[\tau] - \sum_{t=T+1}^{\infty} \mathbb{P}(t < \tau)) \\
&= k - \tfrac{1}{3} \sum_{t=T+1}^{\infty} \mathbb{P}(t < \tau).
\end{aligned}$$

To bound $\sum_{t=T+1}^{\infty} \mathbb{P}(t < \tau)$, Let $N_t = |C_t|$. We have,

$$\begin{aligned}
\mathbb{P}(t < \tau) = \mathbb{P}(N_{t-1} < k+1) &= \mathbb{P}\left(\frac{N_{t-1}}{t-1} < \frac{k+1}{t-1}\right) = \mathbb{P}\left(\frac{N_{t-1}}{t-1} - \mathbb{E}\left[\frac{N_{t-1}}{t-1}\right] < \frac{k+1}{t-1} - \mathbb{E}\left[\frac{N_{t-1}}{t-1}\right]\right) \\
&= \mathbb{P}\left(\frac{N_{t-1}}{t-1} - \mathbb{E}\left[\frac{N_{t-1}}{t-1}\right] < \frac{k+1}{t-1} - \frac{1}{3}\right) < e^{-2(t-1)(\frac{k+1}{t-1} - \frac{1}{3})^2} < e^{-(\frac{2(t-1)}{9} - \frac{4}{3}(k+1))}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{t=T}^{\infty} \mathbb{P}(t < \tau) &< \sum_{t=T}^{\infty} e^{-(\frac{2(t-1)}{9} - \frac{4(k+1)}{3})} = e^{\frac{4(k+1)}{3}} \sum_{t=T}^{\infty} e^{-\frac{2}{9}(t-1)} \\
&\leq e^{\frac{4(k+1)}{3}} \int_{t=T-1}^{\infty} e^{-\frac{2}{9}(t-1)} dt = -\frac{9}{2} e^{\frac{4(k+1)}{3}} \cdot \left[e^{-\frac{2}{9}(t-1)}\right]_{T-1}^{\infty} \\
&= \frac{9}{2} e^{\frac{4(k+1)}{3}} \cdot e^{-\frac{2}{9}(T-1)}.
\end{aligned}$$

Setting $T > 6k + 14$ makes this smaller than 1. Hence $\mathbb{E}[\sum_{t=1}^{T} R_t] \geq k - 1/3$. To verify that τ is defined with a high probability, note that $\mathbb{P}(T < \tau) \leq e^{-(\frac{2(T-1)}{9} - \frac{4}{3}(k+1))} \leq 1/k$. Plugging this into Eq. 3 gives the desired lower bound.

■