Proceedings of Machine Learning Research vol 313:1–32, 2026

Discriminative Feature Feedback with General Teacher Classes

Omri Bar Oz BAROZOM@POST.BGU.AC.IL
Faculty of Computer and Information Science, Ben-Gurion University of the Negev

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

Abstract

We study the theoretical properties of the interactive learning protocol *Discriminative Feature Feedback* (DFF) (Dasgupta et al., 2018). The DFF learning protocol uses feedback in the form of *discriminative feature explanations*. We provide the first systematic study of DFF in a general framework that is comparable to that of classical protocols such as supervised learning and online learning. We study the optimal mistake bound of DFF in the realizable and the non-realizable settings, and obtain novel structural results, as well as insights into the differences between Online Learning and settings with richer feedback such as DFF. We characterize the mistake bound in the realizable setting using a new notion of dimension. In the non-realizable setting, we provide a mistake upper bound and show that it cannot be improved in general. Our results show that unlike Online Learning, in DFF the realizable dimension is insufficient to characterize the optimal non-realizable mistake bound or the existence of no-regret algorithms.

Keywords: discriminative feature feedback, interactive learning, mistake bound, teacher class

1. Introduction

In this work, we study the theoretical properties of the interactive learning protocol *Discriminative Feature Feedback* (DFF) (Dasgupta et al., 2018). Classical learning protocols, such as Supervised Learning (Vapnik and Chervonenkis, 1971), Online Learning (Littlestone, 1988) and Active Learning (McCallum and Nigam, 1998), involve interaction between the learner and the environment that is based only on examples and labels. However, learning in real-world environments can involve many other forms of interactions that provide rich feedback and can speed up learning. The theoretical study of protocols with rich feedback can shed light on their properties.

The DFF learning protocol uses feedback in the form of *discriminative feature explanations*. In this setting, the learner provides an example as an explanation for each of its predictions, and the teacher provides the correct label and a feature explanation whenever the learner's prediction is incorrect. As an illustrative example, suppose the learner classifies a given image of an animal as a zebra, and provides as explanation an image of a previously seen zebra. Then, the teacher responds that the animal is actually a horse, and explains that unlike the example provided by the learner, the animal in the new image does not have stripes. Thus, the provided feature discriminates between the two examples and explains why they have different labels. This type of feedback has been successfully used in practical applications (e.g., Branson et al., 2010; Zou et al., 2015; Liang et al., 2020).

The theory of DFF has been studied so far (Dasgupta et al., 2018; Dasgupta and Sabato, 2020; Sabato, 2023) under a specific component model that assumes that the domain of examples is covered by a small number of subsets of examples that have the same label, and that each pair of subsets with different labels can be discriminated by a single feature. Mistake upper bounds and lower bounds have been obtained for this model. However, DFF has not been studied so far in more generality.

In this work, we provide the first systematic study of DFF in a general framework that allows any class of teachers. Here, a teacher defines both the true labeling function and the features that would be provided for each possible pair of a new example and a previous example that is provided as an explanation. Studying classes of teachers provides DFF with a framework comparable to that of classical protocols such as Supervised Learning and Online Learning. We study the optimal mistake bound of DFF in the realizable and the non-realizable settings, and obtain novel structural results, as well as insights into the difference between Online Learning and settings with richer feedback such as DFF.

Contributions We define the notion of a general teacher class for DFF, and present a new relevant dimension of a teacher class, called DFFdim (Section 4). We show that this dimension characterizes the optimal mistake bound of a teacher class in the realizable setting (Theorem 5), using a new Standard Optimal Algorithm for DFF. We demonstrate the use of the DFFdim by analyzing a less restrictive version of the component model of Dasgupta et al. (2018). We then study the relationship between Online Learning and DFF (Section 5), by providing a two-way mapping between the two settings that preserves important properties (Theorem 8, Theorem 9) and allows directly comparing the mistake bounds of pairs of mapped problems. We use this to show a strong separation between Online Learning and DFF: We show a teacher class with a DFFdim of 1, while the Littlestone dimension of its Online Learning counterpart is infinite (Theorem 10).

Lastly, we study the *non-realizable setting*. We provide a mistake upper bound (Theorem 11) via a simple standard algorithm that employs an optimal algorithm for the realizable setting as a subroutine. This algorithm is quite general, and provides a mistake upper bound for a wide range of natural interactive protocols (Theorem 12). We further show that this upper bound cannot be improved for DFF with general teacher classes (Theorem 13). To prove this lower bound, we use the idea of a *secret-sharing scheme*, a well-known tool used in cryptography (Shamir, 1979). We construct a teacher class using a set of such schemes that depend on the provided explanation. From this lower bound, we conclude that there are no general no-regret algorithms for DFF problems with a finite DFF dimension. Nonetheless, we show that no-regret algorithms do exist for some DFF teacher classes, and conclude that unlike Online Learning, the optimal non-realizable mistake bound cannot be fully characterized using the realizable dimension. This raises an intriguing open question regarding the relationship between the properties of an interactive learning problem and its tolerance to teacher mistakes.

Our analysis and results for the DFF setting provide a glimpse into the intricate landscape of learning using rich interactive learning protocols, which we hope will inspire further research.

2. Related work

Interactive learning protocols are widely used in practical applications (see, e.g., Mosqueira-Rey et al., 2022). For instance, Teso and Kersting (2019) use machine-generated explanations and corrections by users. Feature feedback was proposed as an aid for learning as early as Croft and Das (1990). Similar ideas have been applied in various applications (e.g., Raghavan et al., 2005; Druck et al., 2008; Settles, 2011; Mac Aodha et al., 2018). Learning with explanations has also been studied for neural networks (Schramowski et al., 2020) and Large Language Models (Lampinen et al., 2022).

The formal DFF protocol was first defined in Dasgupta et al. (2018). They further defined the component model and showed an algorithm and a mistake bound for this model in the realizable setting. Subsequently, Dasgupta and Sabato (2020) introduced a non-realizable version of DFF, in which the teacher might not always adhere to the protocol. They presented a robust algorithm for the component model, along with a mistake upper bound that depends on the number of rounds in which the teacher deviates from the protocol. An improved algorithm for the non-realizable component model was presented in Sabato (2023). Other types of feature feedback have been theoretically studied in Poulis and Dasgupta (2017); Visotsky et al. (2019). More generally, Hanneke et al. (2022) study an interactive setting in which the learner can make arbitrary binary-valued queries, and Yadav et al. (2024) study auditing with explanations.

We note that the term "teacher" is also used in the Machine Teaching literature (see, e.g., Zilles et al., 2008; Doliwa et al., 2014). However, our use of this term is unrelated to the Machine Teaching paradigm.

3. Preliminaries

For an integer n, denote $[n] = \{1, \ldots, n\}$. We assume a domain of examples $\mathcal{X}$ and a finite domain of labels $\mathcal{Y}$. In addition, we consider Boolean *features*, which are defined as functions $\phi : \mathcal{X} \to \{\texttt{true}, \texttt{false}\}$. For convenience, we sometimes use 1 to mean `true` and 0 to mean `false`. Thus, we may say that $\phi(x)$ holds or that $\phi(x) = 1$. Denote the negation of ϕ by $\neg\phi \equiv 1 - \phi$. We denote the set of available Boolean features for a given learning problem by Φ. We generally assume that Φ is closed under negation. In various contexts, we use $\perp$ to represent a null element. We assume $\perp \notin \mathcal{X}, \mathcal{Y}, \Phi$. We use '$\cdot$' to represent an arbitrary value. For instance, $(x, \cdot)$ represents a pair with x as the first element and an arbitrary second element.

The Discriminative Feature Feedback (DFF) protocol (Dasgupta et al., 2018) is defined as follows. At every round t,

- A new instance x_t arrives.
- The learning algorithm provides a predicted label $\hat{y}_t$, and an instance $\hat{x}_t$ that was previously observed with that label. This instance serves as 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 correct, no additional feedback is obtained from the teacher.
- If the prediction is incorrect, the teacher provides the correct label of x_t, denoted y_t, and a feature $\phi \in \Phi$ that explains why x_t does not have the same label as $\hat{x}_t$. ϕ is satisfied by x_t but not by $\hat{x}_t$.

It is further assumed that at least one fully labeled example is provided to the learner before the first round, thus providing at least one possible response for the first prediction round. We call the fully labeled examples given to the learner in advance the *history*, usually denoted $H \subseteq \mathcal{X} \times \mathcal{Y}$.

Dasgupta et al. (2018) studied a specific *component model* for teachers, which assumes that the true teacher is consistent with some unknown cover of the domain. Each subset in the cover is called a *component* and all the examples in a given subset have the same label. It is assumed that the teacher always provides a "discriminative" feature feedback ϕ_t: a feature that is satisfied by all the examples in the component of x and is not satisfied by all the examples in the component of $\hat{x}$. In this work, we study a more general version of DFF, in which we allow any *teacher class*, an analog to the hypothesis classes of standard supervised learning analysis. The teacher class defines a given DFF problem. The component model is one such teacher class.

We first provide a general definition of a *teacher*. In the realizable setting, the true teacher determines the feedback that would be provided to the learner in each round. Thus, the teacher determines both the labeling function and the feature feedback function in any possible interaction with the learner.

Definition 1 (Teacher) *A* teacher T *over* $\mathcal{X}, \mathcal{Y}, \Phi$ *is a pair* (ℓ, ψ)*, where* $\ell : \mathcal{X} \to \mathcal{Y}$ *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* $\ell(x) \neq \ell(\hat{x})$ *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 then defined as 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.

The number of mistakes made by a DFF algorithm $\mathcal{A}$ in a specific run is the number of times its label prediction was incorrect. Given a teacher class $\mathcal{T}$ and a history H, the worst-case mistake bound of $\mathcal{A}$ over all possible input sequences, assuming feedback that is consistent with some teacher in $\mathcal{T}$, is denoted $M(\mathcal{A}, \mathcal{T}, H)$.

Under this framework, the component model of Dasgupta et al. (2018) is a teacher class $\mathcal{T}_m$ for $m \in \mathbb{N}$ that includes all teachers that are consistent with a component model that includes m components. The analyses in Dasgupta et al. (2018); Dasgupta and Sabato (2020); Sabato (2023) provide bounds on $M(\mathcal{A}, \mathcal{T}_m, H)$ that depend on m for specific algorithms $\mathcal{A}$. In this work, we study the optimal mistake bound for general teacher classes.

4. The optimal mistake bound for realizable DFF

The VC-dimension (Vapnik and Chervonenkis, 1971) maps each hypothesis class to an integer that characterizes the sample complexity of the optimal algorithm in the PAC setting. The Littlestone dimension (Littlestone, 1988) provides the optimal mistake bound achievable in the Online Learning setting, using a shattered tree construction as a witness. We define the Discriminative Feature Feedback Dimension (DFFdim) for a given teacher class, using a different tree construction, *DFF tree* (DFFT), that aligns with the DFF protocol. We show that the dimension witnessed by a shattered DFFT characterizes the optimal mistake bound for DFF learning. A main difference between a Littlestone tree and the DFFT that we define is that in a Littlestone tree, the responses of the algorithm do no need to be directly encoded, since they do not affect the teacher's responses. This is not the case in DFF, since a different explanation by the algorithm may lead to a different feature response by the teacher. In particular, this means that unlike a Littlestone tree, in a DFFT multiple paths from the root to a leaf can be consistent with the same teacher.

A DFFT (see Figure 1 for an illustration, and Figure 4 below for a concrete example of a shattered DFFT induced by a specific teacher class.) is a rooted tree that represents options for interaction between the environment and the learner. In a DFFT, each node represents an action by the environment, which includes the feedback from the teacher on the last prediction of the learner, as well as the next example to be presented. The root node specifies only the first example to be presented, without any feedback. Each node is of the form $\langle y, \phi, x\rangle$, where $y \in \mathcal{Y} \cup \{\perp\}$, $\phi \in \Phi \cup \{\perp\}$ and $x \in \mathcal{X} \cup \{\perp\}$. Here, y and ϕ represent the teacher feedback for the latest prediction of the algorithm, and x represents the next example. We use $y = \perp$ if this is the first round and there was no previous prediction. We use $\phi = \perp$ if this is the first round, or if the last prediction of the algorithm was correct. We use $x = \perp$ if this is the last round so no example comes next.

Each edge between a parent node and a child node is labeled by the prediction and the explanation example that the algorithm selected as a response to the example in the parent node. Edge labels are thus pairs $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$, where $\hat{y}$ is the prediction for the example in the parent node and $\hat{x}$ is the algorithm's explanation for this prediction.

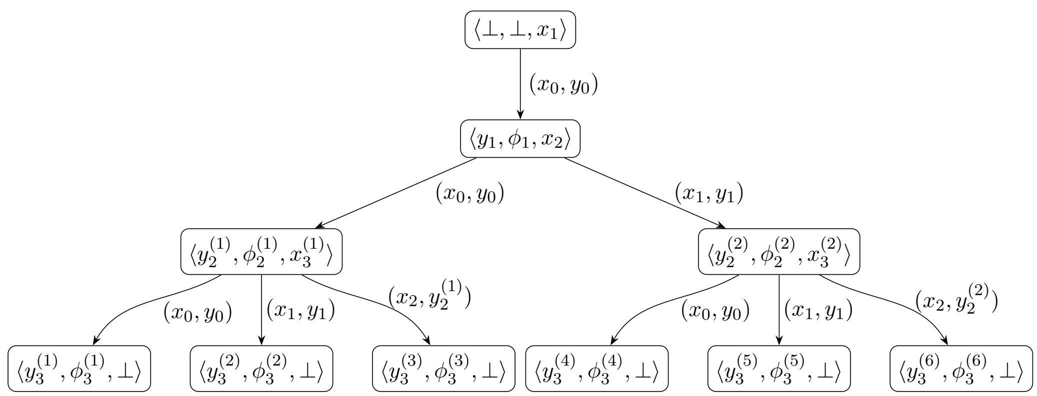

Figure 1: A shattered DFFT with history $H = \{(x_0, y_0)\}$ and height 3.

Definition 2 (DFF tree) *A* DFF tree *(DFFT) over $\mathcal{X}, \mathcal{Y}, \Phi$ is a rooted tree in which the following hold.*

1. *All nodes are of the form $\langle y, \phi, x\rangle$ for $y \in \mathcal{Y} \cup \{\perp\}$, $\phi \in \Phi \cup \{\perp\}$ and $x \in \mathcal{X} \cup \{\perp\}$.*
2. *In the root node, $y = \phi = \perp$.*
3. *A node has $x = \perp$ if and only if the node is a leaf.*
4. *Each edge is labeled by a pair $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$.*
5. *For every non-root node $\langle y, \phi, x\rangle$ with an incoming edge $(\hat{x}, \hat{y})$, if $y \neq \hat{y}$ then $\phi \neq \perp$.*

A DFFT is not necessarily consistent with a given teacher class or history. A *shattered DFFT*, which we define below, will allow us to define the DFF dimension of a given teacher class with a given history. Denote parent and child nodes v_1, v_2 that are connected by an edge labeled e by $v_1 \xrightarrow{e} v_2$. We say that a path from the root to a node in a given DFFT is *consistent* with a given teacher $T = (\ell, \psi)$, if for every pair of parent and child nodes in the path of the form $\langle \cdot, \cdot, x\rangle \xrightarrow{(\hat{x},\hat{y})} \langle y, \phi, \cdot\rangle$, we have $\ell(x) = y$ and if $y \neq \hat{y}$, $\psi(x, \hat{x}) = \phi$. We call every such pair (x, y) on a given path a *labeled example* in the path.

Since a DFF algorithm can only provide examples that were previously observed, the definition of a shattered DFFT takes into account the provided history, which is a non-empty set $H \subseteq \mathcal{X} \times \mathcal{Y}$ of labeled examples. We say that a teacher $T = (\ell, \cdot)$ is consistent with a history H if for all $(x, y) \in H, \ell(x) = y$. We say that a teacher class $\mathcal{T}$ is consistent with a history H (and vice versa) if there is at least one teacher in $\mathcal{T}$ which is consistent with H. Denote by $\mathcal{T}_H$ the set of teachers that are consistent with H in $\mathcal{T}$. We can now define the notion of a shattered DFFT; See illustration in Figure 1.

Definition 3 (Shattered DFFT) *Let $\mathcal{T}$ be a class of teachers over $\mathcal{X}, \mathcal{Y}, \Phi$ and let $H \subseteq \mathcal{X} \times \mathcal{Y}$ be a (non-empty) history that is consistent with $\mathcal{T}$. A DFFT over $\mathcal{X}, \mathcal{Y}, \Phi$ is shattered by $\mathcal{T}$ and H if the following conditions hold.*

1. *For any node $\langle y, \phi, x\rangle$ with an incoming edge $(\hat{x}, \hat{y})$, $y \neq \hat{y}$.*
2. *The outgoing edges of each non-leaf node v are labeled by exactly all the pairs (x, y) that satisfy at least one of the following conditions:*
   - *$(x, y) \in H$, or*
   - *(x, y) is a labeled example in the path from the root of the tree to v.*
3. *Every path from the root to a leaf in the tree is consistent with at least one teacher from $\mathcal{T}_H$.*
4. *The tree is complete; that is, all paths from the root to a leaf are of the same length.*

The height of the tree is the number of edges in any path from the root to a leaf, which is also the number of examples in nodes along this path. We can now define the DFF dimension, DFFdim.

Definition 4 (DFFdim) *Let $\mathcal{T}$ be a teacher class and let $H \subseteq \mathcal{X} \times \mathcal{Y}$ be a (non-empty) history that is consistent with $\mathcal{T}$. $\mathrm{DFFdim}(\mathcal{T}, H)$ is the maximal integer d such that there exists a DFFT of height d that is shattered by $\mathcal{T}$ and H.*

The following theorem shows that the DFFdim of a teacher class $\mathcal{T}$ with a given history H characterizes the optimal mistake bound for an algorithm for these $\mathcal{T}$ and H.

Theorem 5 *Let $\mathcal{T}$ be a teacher class and let $H \subseteq \mathcal{X} \times \mathcal{Y}$ be a (non-empty) history that is consistent with $\mathcal{T}$. Then $\min_{\mathcal{A}} M(\mathcal{A}, \mathcal{T}, H) = \mathrm{DFFdim}(\mathcal{T}, H)$, where the minimum is taken over all deterministic DFF algorithms. In particular, this minimum is attained by the DFF algorithm* `SOA-DFF` *given in Alg. 1.*

`SOA-DFF` (Alg. 1), which attains the optimal realizable mistake bound, is a Standard Optimal Algorithm for DFF. Given a teacher class $\mathcal{T}$, we denote its restriction based on a single round of DFF interaction by $\mathcal{T}|_{(x,\hat{x},y,\phi)} = \{T \in \mathcal{T} \mid T = (\ell, \psi) \wedge \ell(x) = y \wedge \psi(x, \hat{x}) = \phi\}$. Given a teacher class and a history, the algorithm selects in each round, out of its available responses $(\hat{x}, \hat{y})$, the response that in the worst case, would reduce the DFF dimension the most. The future DFF dimension is calculated with respect to the future history, which is updated based on the possible teacher responses. For convenience, we define $\mathrm{DFFdim}(\emptyset, H) = -1$ for any history H. We prove that every time a mistake is made, the DFFdim of the remaining teachers is reduced by at least 1. Theorem 5 is proved in Appendix A.

We note that the DFFdim also characterizes the optimal mistake bound only for deterministic algorithms. This is similar to the Littlestone dimension for Online Learning. In Filmus et al. (2023), a version of the Littlestone dimension for randomized algorithm is proposed. We defer a similar endeavor for DFF to future work.

Example: The DFFdim of a relaxed component model To demonstrate the usefulness of the DFFdim, we analyze a relaxed version of the component model that was studied in Dasgupta et al. (2018). In our setting, each component is defined via a conjunction of features that hold for all examples in the component, as well as a common label. Unlike the component model of Dasgupta et al. (2018), the relaxed version does not require the components to cover the domain. Examples that are not covered are assumed to share a "default" label. In addition, the relaxed model does not require the existence of a discriminative feature between pairs of components. In Appendix B, we give a formal definition of the model and show that if there are R components with at most M features in each conjunction, then the DFF dimension is at most RM, and this is tight for a natural maximal case. This result provides an optimal mistake bound for this problem.

In the next section, we discuss the relationship between DFF and Online Learning.

5. Converting between DFF and Online Learning

What is the relationship between Online Learning and DFF? Intuitively, DFF is Online Learning with the potential for a reduced mistake bound due to feature feedback. We now formalize this correspondence and use it to derive a separation between the two settings. First, we define a mapping from Online Learning problems to DFF problems and vice versa. An Online Learning problem can be converted to an equivalent DFF problem by providing unhelpful feature feedback. In the other direction, a DFF problem can be converted to an Online Learning problem, but the result may have a higher mistake bound. The relationship in both directions is shown by comparing the Littlestone dimension of the Online Learning problem to the DFF dimension of the DFF problem. We then show an example in which the Online Learning version has an infinite Littlestone dimension, while its DFF counterpart has a DFF dimension of 1.

We define the mapping OtD, which maps an Online Learning problem given by a hypothesis class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ to a DFF problem given by a teacher class $\mathcal{T}$ and history H. The mapping defines

Algorithm 1 Standard Optimal Algorithm for Discriminative Feature Feedback (`SOA-DFF`)

procedure `SOA-DFF`($\mathcal{T}, H$)
  $V^1 \leftarrow \mathcal{T}$, $H^1 \leftarrow H$
  for $t = 1, 2, \ldots$ **do**
    Receive x_t
    $(\hat{x}_t, \hat{y}_t) \leftarrow \mathrm{argmin}_{(\hat{x},\hat{y}) \in H^t} \max_{y \neq \hat{y}, \phi \in \Phi} \mathrm{DFFdim}(V^t|_{(x_t, \hat{x}, y, \phi)}, H^t \cup \{(x_t, y)\})$.
    Predict $\hat{y}_t$ and provide $\hat{x}_t$ as an explanation.
    Receive feedback y_t and ϕ_t.
    $V^{t+1} \leftarrow V^t|_{(x_t, \hat{x}_t, y_t, \phi_t)}$, $H^{t+1} \leftarrow H^t \cup \{(x_t, y_t)\}$

explanations that do not divulge any additional information over the provided label. We further provide the converse mapping DtO, and show that this recovers the Online Learning problem back from a mapped DFF problem. Moreover, any DFF algorithm for the DFF problem $\mathrm{OtD}(\mathcal{F})$ can be converted to an equivalent Online Learning algorithm that obtains the same number of mistakes, by simply ignoring the explanations of the teacher and refraining from providing explanation examples. We show that OtD preserves the optimal mistake bound of the problem by comparing the respective dimensions. For a set of labeled examples $S \subseteq \mathcal{Y}^{\mathcal{X}}$, denote $S_{\mathcal{X}} := \{x \mid \exists y \in \mathcal{Y} \text{ s.t. } (x, y) \in S\}$. We two mappings are defined as follows.

Definition 6 (Mapping from Online Learning to DFF) *Given a hypothesis class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ defining an Online Learning problem, for any $y \in \mathcal{Y}$ let $\star_y$ denote a special example that is not in $\mathcal{X}$. Let $H := \{(\star_y, y) \mid y \in \mathcal{Y}\}$, and define $\mathcal{X}' := \mathcal{X} \cup H_{\mathcal{X}}$. Let the set of features over $\mathcal{X}'$ be $\Phi := \{\mathbb{I}[x] \mid x \in \mathcal{X}'\}$, where $\mathbb{I}[x](x') = 1 \iff x = x'$.*

For any $f \in \mathcal{F}$, define $f' : \mathcal{X}' \to \mathcal{Y}$ by $f' := \{(x, f(x)) \mid x \in \mathcal{X}\} \cup \{(\star_y, y) \mid y \in \mathcal{Y}\}$. Let $\psi_f : \mathcal{X}' \times \mathcal{X}' \to \Phi \cup \{\bot\}$ be a feature feedback function defined as $\psi_f(x, x') = \mathbb{I}[x]$. Let $T_f := (f', \psi_f)$ be a teacher over $\mathcal{X}', \mathcal{Y}, \Phi$. Define the mapping of $\mathcal{F}$ from Online Learning to DFF by $\mathrm{OtD}(\mathcal{F}) := (\mathcal{T}_{\mathcal{F}}, H)$, where $\mathcal{T}_{\mathcal{F}} = \{T_f \mid f \in \mathcal{F}\}$.

The converse mapping, DtO, is defined next.

Definition 7 (Mapping from DFF to Online Learning) *Given a DFF problem defined by a teacher class $\mathcal{T}$ over $\mathcal{X}, \mathcal{Y}, \Phi$ and a history $H \subseteq \mathcal{X} \times \mathcal{Y}$, let $\mathcal{X}' = \mathcal{X} \setminus H_{\mathcal{X}}$. For a teacher $T = (\ell, \psi) \in \mathcal{T}$, let $f_T := \ell|_{\mathcal{X}'}$. The mapping of the DFF problem $(\mathcal{T}, H)$ to Online Learning is $\mathrm{DtO}(\mathcal{T}, H) = \mathcal{F}$, where $\mathcal{F} = \{f_T \mid T \in \mathcal{T}\}$.*

First, we show that mapping from Online Learning to DFF and back recovers the original online learning problem, thus proving that no information is lost in the mapping. The proof is provided in Appendix C.

Theorem 8 *For any hypothesis class $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$, we have $\mathrm{DtO}(\mathrm{OtD}(\mathcal{F})) = \mathcal{F}$.*

The next theorem shows that the Littlestone dimension of an Online Learning problem is equal to the DFF dimension of its mapping to a DFF problem. In other words, the optimal mistake bound of the problem is preserved by the conversion. Denote the Littlestone dimension of an Online Learning problem $\mathcal{F}$ by $\mathrm{Ldim}(\mathcal{F})$. Note that since our problems can have more than two labels, we use the multiclass Littlestone dimension, which is also witnessed by a binary tree (Daniely et al., 2015). The theorem is proved in Appendix C. We give here a partial sketch.

Theorem 9 *For any hypothesis class $\mathcal{F} \in \mathcal{Y}^{\mathcal{X}}$, $\mathrm{DFFdim}(\mathrm{OtD}(\mathcal{F})) = \mathrm{Ldim}(\mathcal{F})$.*

Proof [Sketch for the lower bound] Given a shattered Littlestone tree for $\mathcal{F}$ with height $\mathrm{Ldim}(\mathcal{F})$, we construct $\mathtt{Tr}'$, a shattered DFFT of the same height, by inductively mapping from nodes in $\mathtt{Tr}$ to nodes in $\mathtt{Tr}'$. For a node v in $\mathtt{Tr}'$, $N(v)$ denotes that the node in $\mathtt{Tr}$ that it was mapped from. Let y_1, y_2 be the two outgoing edges from $N(v)$ in $\mathtt{Tr}$, and let x_1, x_2 be the respective examples in the target nodes u_1, u_2 of these edges in $\mathtt{Tr}$ (see Figure 2). In $\mathtt{Tr}'$, we set the target of each outgoing edge $(\hat{x}, \hat{y})$ to a node $\langle y_1, \mathbb{I}[x], x_1 \rangle$ if $\hat{y} \neq y_1$, and to $\langle y_2, \mathbb{I}[x], x_2 \rangle$ if $\hat{y} = y_1$. We prove that this inductive construction results in a shattered DFFT. ■

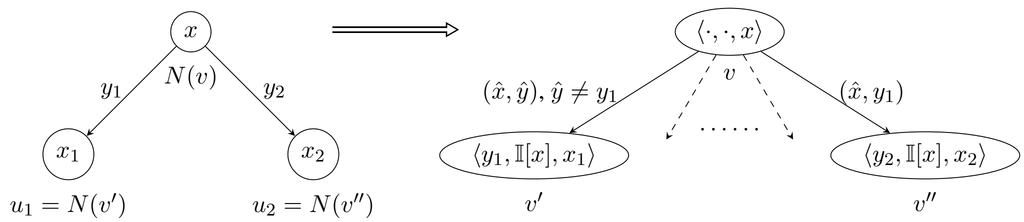

Figure 2: Illustrating the construction of $\mathtt{Tr}'$ in the proof of Theorem 9

Using the mappings defined above, we can compare the optimal mistake bound of a DFF problem with the one of its Online Learning equivalent, in which no explanations are provided. A strong separation between the two settings is demonstrated by the following example, in which the DFFdim is 1, while the Littlestone Dimension is infinite. Moreover, this example has an infinite Littlestone tree, implying that an adversary can force the algorithm to make a mistake in every round in an infinite sequence of rounds (see Bousquet et al., 2021). In Dasgupta et al. (2018), the power of DFF compared to standard Online Learning was discussed in the context of computational complexity and CNF formulas. The example and theorem below provide a clear statistical separation.

We use the following *natural feature construction*: Let $\mathcal{X} = \{0,1\}^{\mathbb{N}}$, and define the Boolean function f_n by $f_n(x) := x(n)$. Set $\Phi := \cup_{n\in\mathbb{N}}\{f_n, \neg f_n\}$. F'or vectors in $\{0,1\}^{\mathbb{N}}$, we use an overline to indicate infinite repetition, so that $\bar{0}, \bar{1}$ are the all-zero and all-one vectors, and $\overline{01}$ is a vector of alternating coordinate values.

Example 1 *Let $\mathcal{Y} = \{0,1\}$ and let $\mathcal{X}, \Phi$ be the natural feature construction defined above. For any $n \in \mathbb{N}$, define a teacher $T_n = (f_n, \psi_n)$, where*

$$\psi_n(x, x') := \begin{cases} f_n & f_n(x) = 1, f_n(x') = 0 \\ \neg f_n & f_n(x) = 0, f_n(x') = 1 \\ \perp & \textit{otherwise} \end{cases}$$

Let the teacher class be $\mathcal{T} = \{T_n \mid n \in \mathbb{N}\}$. Define $H = \{(\bar{0}, 0), (\bar{1}, 1)\}$.

Theorem 10 *For $\mathcal{T}, H$ as defined in Example 1, Then $\mathrm{DFFdim}(\mathcal{T}, H) = 1$ while $\mathrm{Ldim}(\mathrm{DtO}(\mathcal{T}, H)) = \infty$. Moreover, there exists an infinite Littlestone tree for $\mathrm{DtO}(\mathcal{T}, H)$.*

Proof Let $\mathcal{X}' = \mathcal{X} \setminus \{\bar{0}, \bar{1}\}$. Let $\mathcal{F} := \mathrm{DtO}(\mathcal{T}, H)$. Then $\mathcal{F} = \{f_n|_{\mathcal{X}'} \mid n \in \mathbb{N}\}$. To see that $\mathrm{Ldim}(\mathcal{F}) = \infty$, consider a complete binary Littlestone tree constructed as follows: Set the root node to $x = \overline{01}$. The rest of the tree is constructed inductively, as follows. For a given node v, let $(x_1, y_1), \ldots, (x_t, y_t)$ be the sequence of labeled examples in the path from the root to v. Let $A_v := \{n \mid \forall i \in [t], x_i(n) = y_i\}$. Set the example in node v to some x such that $x(A_v) = \overline{01}$. It can be seen by induction that for each v, $|A_v| = \infty$. Therefore, for any path from the root to a leaf v, any f_n for $n \in A_v$ is consistent with the labeled examples on the path. It follows that there exists a Littlestone tree of any finite size for $\mathcal{F}$, as well as a Littlestone tree of infinite depth.

To show that $\mathrm{DFFdim}(\mathcal{T}, H) = 1$, it suffices to provide a deterministic algorithm that makes at most a single mistake, since clearly no algorithm for this problem has a mistake upper bound of 0.

Consider the algorithm that always predicts the label 0 with the explanation $\bar{0}$, until it makes the first mistake on some example. The true label of the example is 1, and the feature feedback provided by the teacher for this example is the true labeling function. Thereafter, the algorithm predicts using the true labeling function and does not make any more mistakes. This completes the proof. ■

6. DFF in the non-realizable setting

The previous sections discuss DFF in the realizable setting, in which the feedback provided to the algorithm is assumed to be consistent with one of the teachers in the teacher class. Previous works (Dasgupta and Sabato, 2020; Sabato, 2023) studied a non-realizable setting for the original component model of Dasgupta et al. (2018), and provided mistake upper bounds. In the non-realizable setting, it is assumed that the feedback provided to the algorithm is consistent with one of the teachers in the teacher class, except for at most $k \in \mathbb{N}$ rounds, in which the feedback might have a different label and/or a different feature feedback than prescribed by the teacher. This behavior is called an *exception* from the protocol. Rounds with an exception are called *exception rounds*. We call the setting with up to k exception rounds the k*-non-realizable* setting. Here, we provide new bounds for general teacher classes for this setting.

In Online Learning, the k-non-realizable setting is equivalent to assuming that at most k of the labels provided by the environment are different from those of the true labeling function. For an Online Learning problem with a hypothesis class $\mathcal{F}$, the optimal mistake bound assuming at most k label errors is $k + \Theta\left(\sqrt{k \cdot \mathrm{Ldim}(\mathcal{F})} + \mathrm{Ldim}(\mathcal{F})\right)$, where the upper bound is obtained by a randomized algorithm (Filmus et al. 2023, following Ben-David et al. 2009; Alon et al. 2021).

The apparent similarity of DFF to Online Learning, as discussed in Section 5, may lead to the expectation that the dependence on k in mistake bounds for DFF would be similar to that of Online Learning. However, we show that this is not the case. We obtain a mistake upper bound of $O(k \cdot \mathrm{DFFdim}(\mathcal{T}, H))$, using a simple deterministic algorithm for the k-non-realizable setting. We then show that this upper bound cannot be improved for general DFF problems, even with a randomized algorithm, assuming a mild version of adversarial adaptivity that holds also for the Online Learning bound mentioned above. We conclude that unlike Online Learning, a no-regret algorithm for general DFF problems with a finite dimension does not exist in this setting. Moreover, again unlike Online Learning, the dependence of the optimal mistake bound on k is highly dependent on the DFF problem, raising a new open question regarding the factors that control this dependence.

The mistake upper bound for the k-realizable setting is obtained using the simple Agnostic Standard Optimal Algorithm `A-SOA-DFF` listed in Alg. 2. This algorithm runs the `SOA-DFF` algorithm that is meant for the realizable setting, and restarts it whenever it makes more mistakes than the realizable mistake upper bound. Denote by $M_k^L(\mathcal{A}, \mathcal{T}, H)$ the worst-case number of mistakes made by a DFF algorithm $\mathcal{A}$ for $\mathcal{T}, H$ on a sequence of L rounds in the k-non-realizable setting.

Theorem 11 *For any teacher class $\mathcal{T}$ and history H, and for any $L \in \mathbb{N}$, if $\mathrm{DFFdim}(\mathcal{T}, H) = d$ then*

$$M_k^L(\texttt{A-SOA-DFF}, \mathcal{T}, H) \leq (k+1)d + k.$$

Proof Call each sub-sequence of the run between initializations of `SOA-DFF` a "run segment". In each run segment, `SOA-DFF` makes $d + 1$ mistakes. Let N be the total number of run segments.

Algorithm 2 Agnostic Standard Optimal Algorithm for DFF (A-SOA-DFF)

procedure A-SOA-DFF($\mathcal{T}, H$)
  Initialize SOA-DFF($\mathcal{T}, H$) and set $M \leftarrow 0$.
  for $t = 1, 2, \ldots$ **do**
    Receive x_t and provide it to SOA-DFF; Get output $(\hat{x}_t, \hat{y}_t)$ from SOA-DFF.
    Predict $\hat{y}_t$ and provide $\hat{x}_t$ as an explanation.
    Receive feedback y_t, ϕ_t and provide it to SOA-DFF.
    If $\hat{y}_t \neq y_t$, $M \leftarrow M + 1$.
    If $M = \mathrm{DFFdim}(\mathcal{T}, H) + 1$, re-initialize SOA-DFF($\mathcal{T}, H$) and set $M \leftarrow 0$

Then the total number of prediction mistakes made by the algorithm is at most $N(d+1)+d$. Since the realizable mistake bound of SOA-DFF is $M(\texttt{SOA-DFF}, \mathcal{T}, H) = d$, all teachers in $\mathcal{T}_H$ are inconsistent with all of the N run segments. Therefore, any such teacher is inconsistent with at least N rounds in the whole run. If there are at most k exceptions during the run, it follows that $N \leq k$. Thus, $M_k^L(\texttt{A-SOA-DFF}, \mathcal{T}, H) \leq k(d+1) + d = (k+1)d + k$. ■

We note that the upper bound in fact holds for a wide range of interactive prediction protocols, of which DFF is but one example. We provide a formal statement and proof of the following result in Appendix D.

Theorem 12 (Informally) *For any "natural" interactive protocol with a realizable mistake upper bound of* d*, there is an algorithm that obtains a mistake upper bound of at most* $(k+1)d + k$ *for any run with at most* k *exception rounds.*

We further show that the upper bound of Theorem 11 is tight, by proving a nearly matching lower bound. The lower bound holds also for randomized algorithms with a mildly adaptive adversary. In general online prediction problems, an oblivious adversary fixes the sequence of loss functions in advance, while an adaptive adversary decides on the loss function in each round based on the algorithm's actions in the previous rounds (Arora et al., 2012). However, in Online Learning, both of these adversaries obtain the same optimal mistake bound of $k + \Theta\left(\sqrt{k \cdot \mathrm{Ldim}(\mathcal{F})} + \mathrm{Ldim}(\mathcal{F})\right)$, using the same randomized Weighted Majority algorithm of Littlestone and Warmuth (1994). In the lower bound below, we consider a DFF version of the adaptive adversary, which fixes the label and the feature feedback function ahead of each round based on the algorithm's actions in previous rounds. The effect of other types of adversaries on the optimal mistake bound of DFF is an intriguing question that we leave for future work. Below, $M_k^L(\cdot,\cdot,\cdot)$ denotes the *expected* mistake bound of a randomized algorithm, under a worst-case adaptive adversary.

Theorem 13 *Let* $\mathcal{X} = \mathbb{N}$, $\mathcal{Y} = \{0, 1\}$, $\Phi = \{0,1\}^{\mathcal{X}}$, $H_\bullet = \{(2,0),(1,1)\}$. *Let* $d \geq 2$. *There is a teacher class* $\mathcal{T}^d$ *such that* $\mathrm{DFFdim}(\mathcal{T}^d, H_\bullet) \leq d$ *and for any* $k \geq 1$ *and* $L \geq 4(k+1)d$*, for any algorithm* $\mathcal{A}$ *(including randomized algorithms)*

$$M_k^L(\mathcal{A}, \mathcal{T}^d, H_\bullet) \geq (k+1)d - k - 2.$$

Taken together with Theorem 11, we conclude that the optimal mistake bound for a general teacher class with a DFF dimension of d in the k-non-realizable setting is $(k+1)d \pm \Theta(k)$. In addition, this implies that no-regret algorithms for the general DFF problem do not exist for $d \geq 3$: for $k = L/(4d) - 1$, we obtain a regret bound of $M_k^L(\mathcal{A}, \mathcal{T}^d, H_\bullet) - k \geq (k+1)d - 2k - 2 = \Omega(L)$.

While Theorem 13 shows that the mistake upper bound of Theorem 11 is tight, it does not show a lower bound that holds for all DFF problems. We observe that unlike Online learning, in DFF the dimension that characterizes the realizable mistake bound is insufficient to characterize the mistake bound in the non-realizable setting. To demonstrate this, consider an Online Learning problem $\mathcal{F} \subseteq \mathcal{Y}^\mathcal{X}$ with $\mathrm{Ldim}(\mathcal{F}) = d$ and its DFF counterpart $(\mathcal{T}, H) := \mathrm{OtD}(\mathcal{F})$. By Theorem 9, $\mathrm{DFFdim}(\mathcal{T}, H) = d$. The optimal mistake bound for the DFF problem is the same as that of is Online Learning counterpart, $k + \Theta(\sqrt{kd} + d) \ll \Theta(kd)$, implying also the existence of a no-regret algorithm. As a different example, consider the component model of Dasgupta et al. (2018). They show that the realizable mistake bound for this model with m components is $O(m^2)$, and this is shown to be tight in Dasgupta and Sabato (2020). It follows that the DFFdim of this problem is $d = \Theta(m^2)$. On the other hand, there exists an algorithm for this problem in the k-non-realizable setting with a mistake bound of $O(m^2 + mk) = O(d + \sqrt{d}k)$ (Sabato, 2023), giving yet a different dependence on k. However, no matching lower bound has been shown. These examples raise an interesting open question: which properties of a DFF problem determine the dependence of its optimal k-non-realizable mistake bound on k.

The proof of Theorem 13, provided in Appendix E, employs a teacher class that satisfies the following: If the algorithm provides the same 1-labeled example as an explanation on d different rounds, for examples whose true label is 0, then it obtains d feature feedbacks, that together perfectly identify the true labeling function. However, it is impossible to obtain *any* information on the true labeling function (beyond the information provided by the label feedback) with fewer than d feature feedbacks that were provided for the same explanation. Moreover, it is impossible to obtain such information by combining feature feedbacks provided for different explanations. In addition, the environment can provide feedback in a way that makes it impossible to distinguish between exception and non-exception rounds. In this problem, the environment can cause the algorithm to repeatedly use examples provided in exception rounds as explanations. These examples are actually labeled 0, but are provided to the algorithm with label 1 in an exception round, thus the provided feature feedback when they are used as explanations is uninformative. However, the algorithm cannot identify this before d mistakes have already been made for each of these k explanations.

A core property of the construction above is the ability to perfectly describe the true function using d items of information, while providing no information if there are only $d - 1$ items. We achieve this via secret sharing (Shamir, 1979), an idea commonly used in cryptography. A secret-sharing scheme allows representing a secret using a number of parts, such that only if a sufficient number of the parts is known, the secret is revealed, and no information about the secret is revealed from fewer parts.

More specifically, we employ a (d, n)-*threshold secret sharing scheme* (Shamir, 1979). This is a mapping from a secret s to a set of n partial secrets $\bar{s}$ such that having access to any subset of $\bar{s}$ of size at least d allows one to perfectly reconstruct s, and having access to only a subset of $\bar{s}$ of size $d-1$ or less does not reveal *any* information about the secret s. This scheme is based on Langrange's interpolation theorem (Lagrange, 1795). The latter states that any polynomial P of degree $d - 1$ can be interpolated when having access to d distinct evaluations $((x_1, P(x_1)), \ldots, (x_d, P(x_d))))$ of the polynomial. Moreover, for any d distinct evaluations, there exists a polynomial P that satisfies them. The full scheme is specified in Appendix E. We use this to construct the class $\mathcal{T}^d$ for Theorem

13, in which each possible explanation provides parts of a different secret-sharing scheme. The construction and the full proof of Theorem 13 are provided in Appendix E.

7. Conclusion

In this work, we study Discriminative Feature Feedback in a general framework based on teacher classes, and show new results characterizing its mistake bound in the realizable setting. We further showed separation between DFF and Online Learning in the realizable case. Lastly, we showed that unlike Online Learning, in DFF the optimal mistake bounds in the non-realizable setting exhibit a rich behavior that depends on the specific teacher class. This presents an intriguing open question regarding the interplay between rich feedback and robustness to deviations from the protocol, as well as the challenge of characterizing this behavior based on properties of the learning problem.

Acknowledgments

We thank our colleagues and funding agencies; 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 supported by the Israeli Council for Higher Education (CHE) via the Data Science Research Center at Ben-Gurion University.

References

Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd annual ACM SIGACT symposium on theory of computing*, pages 447–455, 2021.

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 Conference on International Conference on Machine Learning*, pages 1747–1754, 2012.

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.

S. Branson, C. Wah, B. Babenko, F. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *European Conference on Computer Vision*, 2010.

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.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16:2377–2404, 2015.

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.

Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, vc-dimension and sample compression. *The Journal of Machine Learning Research*, 15(1): 3107–3131, 2014.

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.

Aditya Gangrade, Anil Kag, Ashok Cutkosky, and Venkatesh Saligrama. Online selective classification with limited feedback. *Advances in Neural Information Processing Systems*, 34:14529–14541, 2021.

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

Joseph-Louis Lagrange. *Leçons élémentaires sur les mathématiques*. 1795.

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.

Weixin Liang, James Zou, and Zhou Yu. Alice: Active learning with contrastive natural language explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4380–4391, 2020.

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

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

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.

A. K. McCallum and K. Nigam. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, pages 350–358, 1998.

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.

S. Poulis and S. Dasgupta. Learning with feature feedback. In *Twentieth International Conference on Artificial Intelligence and Statistics*, 2017.

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.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Adi Shamir. How to share a secret. *Commun. ACM*, 22(11):612–613, 1979. doi: 10.1145/359168.359176. URL https://doi.org/10.1145/359168.359176.

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.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

Roman Visotsky, Yuval Atzmon, and Gal Chechik. Learning with per-sample side information. In *AGI*, 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.

Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Teaching dimensions based on cooperative learning. In *COLT*, pages 135–146, 2008.

J.Y. Zou, K. Chaudhuri, and A. T. Kalai. Crowdsourcing feature discovery via adaptively chosen comparisons. In *Conference on Human Computation and Crowdsourcing (HCOMP)*, 2015.

Appendix A. Proof of Theorem 5: optimal realizable mistake bound for DFF

Theorem 5 is proved via the following two lemmas.

Lemma 14 *Let $\mathcal{T}, \mathcal{H}$ as in Theorem 5. There exists a DFF algorithm* SOA-DFF *such that* $M(\textit{SOA-DFF}, \mathcal{T}, H) \leq \mathrm{DFFdim}(\mathcal{T}, \mathcal{H})$.

Lemma 15 *Let $\mathcal{T}, \mathcal{H}$ as in Theorem 5. For any deterministic DFF algorithm $\mathcal{A}$,* $M(\mathcal{A}, \mathcal{T}, H) \geq \mathrm{DFFdim}(\mathcal{T}, H)$.

To prove Lemma 14, we first prove a structural property of the DFF dimension.

Lemma 16 *Let $\mathcal{T}_1, \mathcal{T}_2$ be teacher classes such that $\mathcal{T}_1 \subseteq \mathcal{T}_2$. Let $H_1, H_2 \subseteq \mathcal{X} \times \mathcal{Y}$ be (non-empty) histories such that $H_2 \subseteq H_1$. Then* $\mathrm{DFFdim}(\mathcal{T}_1, H_1) \leq \mathrm{DFFdim}(\mathcal{T}_2, H_2)$.

Proof Let $\mathtt{Tr}_1$ be a shattered DFFT for $\mathcal{T}_1, H_1$ of height $\mathrm{DFFdim}(\mathcal{T}_1, H_1)$. We construct a shattered DFFT for $\mathcal{T}_2, H_2$ of the same height, thus proving that $\mathrm{DFFdim}(\mathcal{T}_2, H_2) \geq \mathrm{DFFdim}(\mathcal{T}_1, H_1)$.

$\mathtt{Tr}_2$ is obtained by removing from $\mathtt{Tr}_1$ all edges (and the sub-trees descended from them) that are labeled by $(\hat{x}, \hat{y}) \in H_1 \setminus H_2$, such that $(\hat{x}, \hat{y})$ is not a labeled example in the path from the root of the tree to the source node of the edge. We show that $\mathtt{Tr}_2$ is a shattered DFFT for $\mathcal{T}_2, H_2$ and that it has the same height as $\mathtt{Tr}_1$.

It is easy to see that property 1 (different labels) holds for $\mathtt{Tr}_2$, since it is a sub-graph of $\mathtt{Tr}_1$. Property 2 (outgoing edges) follows by the fact that all remaining outgoing edges of each node satisfy one of the conditions of this property, since $H_2 \subseteq H_1$. In addition, all required edges exist in $\mathtt{Tr}_2$, since they all exist in $\mathtt{Tr}_1$ and are not removed. Property 3 (consistency with a teacher in $(\mathcal{T}_2)_{H_2}$) follows from the same property in $\mathtt{Tr}_1$ for $\mathcal{T}_1, H_1$, since every path from the root to a leaf in $\mathtt{Tr}_2$ is also such a path in $\mathtt{Tr}_1$, and since $\mathcal{T}_2 \supseteq \mathcal{T}_1$ and $H_2 \subseteq H_1$.

Lastly, for property 4 (tree completeness), we show that since $\mathtt{Tr}_1$ is complete, so is $\mathtt{Tr}_2$. Observe that by property 2 of $\mathtt{Tr}_1$ and since $\emptyset \neq H_2 \subseteq H_1$, every node in $\mathtt{Tr}_1$ has at least one outgoing edge $(\hat{x}, \hat{y}) \in H_2$. Therefore, this node is a non-leaf in $\mathtt{Tr}_2$ as well. Thus, all remaining paths from root to leaf in $\mathtt{Tr}_2$ are of the same length as they have in $\mathtt{Tr}_1$. It follows that $\mathtt{Tr}_2$ is also complete. This also implies that $\mathtt{Tr}_2$ has the same height as $\mathtt{Tr}_1$.

To conclude, $\mathtt{Tr}_2$ is a shattered DFFT for $\mathcal{T}_2, H_2$ and has height $\mathrm{DFFdim}(\mathcal{T}_1, H_1)$. Therefore, $\mathrm{DFFdim}(\mathcal{T}_2, H_2) \geq \mathrm{DFFdim}(\mathcal{T}_1, H_1)$. ■

We now prove the mistake bound of SOA-DFF.

Proof [Proof of Lemma 14] We show that for every round t in which SOA-DFF makes a mistake,

$$\mathrm{DFFdim}(V^{t+1}, H^{t+1}) < \mathrm{DFFdim}(V^t, H^t). \tag{1}$$

It follows that SOA-DFF makes at most $\mathrm{DFFdim}(V^1, H^1) = \mathrm{DFFdim}(\mathcal{T}, H)$ mistakes.

By Alg. 1, we have $V^{t+1} = V^t|_{(x_t, \hat{x}_t, y_t, \phi_t)}$ and $H^{t+1} = H^t \cup \{(x_t, y_t)\}$. Thus, by Lemma 16, $\mathrm{DFFdim}(V^{t+1}, H^{t+1}) \leq \mathrm{DFFdim}(V^t, H^t)$. It therefore suffices to show that the inequality is strict. Let t be a round in which SOA-DFF makes a mistake. Assume for the sake of contradiction that $\mathrm{DFFdim}(V^{t+1}, H^{t+1}) = \mathrm{DFFdim}(V^t, H^t)$. By Alg. 1,

$$(\hat{x}_t, \hat{y}_t) = \operatorname*{argmin}_{(\hat{x}, \hat{y}) \in H^t} \max_{y \neq \hat{y}, \phi \in \Phi} \mathrm{DFFdim}(V^t|_{(x_t, \hat{x}, y, \phi)}, H^t \cup \{(x_t, y)\}).$$

Let $H^t = \{(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_n, \tilde{y}_n)\}$. Then, for every $i \in [n]$ there exist $\overline{y}_i \in \mathcal{Y}$ and $\phi_i \in \Phi$ such that $\overline{y}_i \neq \tilde{y}_i$ and, denoting $V_i^t := V^t|_{(x_t, \tilde{x}_i, \overline{y}_i, \phi_i)}$, $H_i^t := H^t \cup \{(x_t, \overline{y}_i)\}$, it holds that

$$\mathrm{DFFdim}(V^{t+1}, H^{t+1}) \leq \mathrm{DFFdim}(V_i^t, H_i^t) \leq \mathrm{DFFdim}(V^t, H^t).$$

The first inequality follows from the choice of $(\hat{x}_t, \hat{y}_t)$, and the right inequality follows from Lemma 16. By the assumption, the RHS and the LHS above are equal. Thus,

$$\mathrm{DFFdim}(V^{t+1}, H^{t+1}) = \mathrm{DFFdim}(V_i^t, H_i^t).$$

For $i \in [n]$, let $\mathtt{Tr}_i$ be a shattered DFFT for V_i^t and H_i^t, with height $\mathrm{DFFdim}(V_i^t, H_i^t) = \mathrm{DFFdim}(V^t, H^t)$. We construct a shattered DFFT for V^t with history H^t with height $\mathrm{DFFdim}(V^t, H^t) + 1$, thus contradicting the definition of DFF dimension.

The shattered DFFT, denoted $\mathtt{Tr}$, is constructed as follows (see Figure 3 for an illustration). The root of $\mathtt{Tr}$ is set to $\langle \bot, \bot, x_t \rangle$. Its outgoing edges are set to edges labeled by $(\tilde{x}_i, \tilde{y}_i)$ for all $i \in [n]$. The target node of the edge labeled $(\tilde{x}_i, \tilde{y}_i)$ is set to $v_i := \langle \overline{y}_i, \phi_i, x_r^i \rangle$, where $\overline{y}_i, \phi_i$ are as defined above and x_r^i is the example at the root of $\mathtt{Tr}_i$. The subtree descending from v_i is set to be identical to the subtree descending from the root of $\mathtt{Tr}_i$.

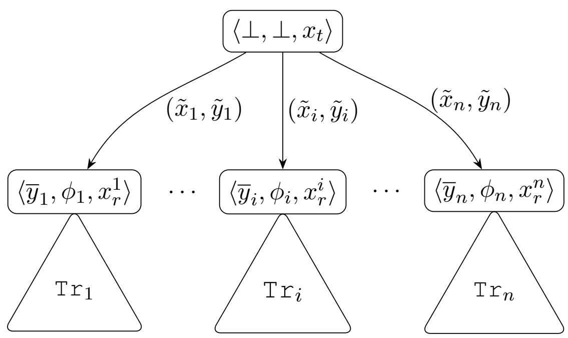

Figure 3: The construction of $\mathtt{Tr}$ in the proof of Lemma 14

It is easy to verify that $\mathtt{Tr}$ is a DFFT of height $\mathrm{DFFdim}(V^t, H^t) + 1$. We now prove that it is shattered by V^t, H^t by verifying the properties in Def. 3. To show property 1 (different labels), let $u = \langle y_u, \phi_u, x_u \rangle$ be a node in $\mathtt{Tr}$, and let its incoming edge be labeled $e = (x_e, y_e)$. Property 1 requires that $y_u \neq y_e$. If the source of this edge is the root node, then $(x_e, y_e) = (\tilde{x}_i, \tilde{y}_i)$ for some i, and $y_u = \overline{y}_i \neq \tilde{y}_i$, satisfying the property. Otherwise, u is a node in one of the $\mathtt{Tr}_i$, in which case the property holds because $\mathtt{Tr}_i$ is shattered.

To show property 2 (edge labels), we first show that all outgoing edges satisfy one of the two possible conditions for labeled edges. Consider an edge e in $\mathtt{Tr}$ labeled by (x_e, y_e) with a source node v. If the edge is outgoing from the root, then by the construction of $\mathtt{Tr}$, $(x_e, y_e) \in H^t$. Otherwise, e is a an edge that in a sub-tree taken from $\mathtt{Tr}_i$. Since $\mathtt{Tr}_i$ is a shattered DFFT, it satisfies property 2, thus one of the two options hold:

- $(x, y) \in H_i^t = H^t \cup \{x_t, \overline{y}_i\}$. If $(x, y) \in H^t$, then the property holds for $\mathtt{Tr}$ using the first case. Otherwise, $(x, y) = (x_t, \overline{y}_i)$. In this case, since the parent-child pair $\langle \bot, \bot, x_t \rangle \to \langle \overline{y}_i, \phi_i, x_r^i \rangle$ are in the path from the root to v, the property is satisfied using the second case.

- In $\texttt{Tr}_i$, these is a labeled example (x, y) in the path from the root to the node that was mapped to v in $\texttt{Tr}$. From the construction, the same holds for v in $\texttt{Tr}$.

Property 2 further requires that the outgoing edges of each non-leaf node are all the possible labeled edges according to the two conditions. For the root node, this holds since the outgoing edges of this node correspond to all the labeled examples in H, and there are no other labeled examples For a node that was mapped from some $\texttt{Tr}_i$ (including its revised root), note that the only difference in the set of possible outgoing edge labels is due to the example $(x_t, \overline{y}_i)$, which is a labeled example that appears in the path to this node in $\texttt{Tr}$ but possibly not in $\texttt{Tr}_i$. The property holds for this node, because it holds for the mapped node in $\texttt{Tr}_i$ with history $H_i^t = H^t \cup \{(x_t, \overline{y}_i)\}$.

Property 3 requires that each path from the root to a leaf is consistent with some teacher in $V^t_{H^t}$. Every such path in $\texttt{Tr}$ is of the form $\langle \perp, \perp, x_t \rangle - (\tilde{x}_i, \tilde{y}_i) \to \langle \overline{y}_i, \phi_i, x_r^i \rangle \to p$, where p is a path in $\texttt{Tr}_i$ from a child of the root to a leaf. Since $\texttt{Tr}_i$ is shattered, there is a teacher from $V_i^t \subseteq V^t_{H_i^t}$, where $H_i^t \supseteq H^t$, that is consistent with the path $\langle \perp, \perp, x_r^i \rangle \to p$. From the definition of H_i^t, it follows that the same teacher is also consistent with the path in $\texttt{Tr}$. Lastly, for Property 4, it is easy to see that since $\texttt{Tr}_i$ are all complete trees, then so is $\texttt{Tr}$.

We have established that $\texttt{Tr}$ is a shattered DFFT for V^t and H^t. It follows that $\mathrm{DFFdim}(V^t, H^t) = 1 + \mathrm{DFFdim}(V_i^t, H_i^t)$ for all i. This implies that $\mathrm{DFFdim}(V^t, H^t) = \mathrm{DFFdim}(V^{t+1}, H^{t+1}) + 1$, in contradiction to Lemma 16. Thus, the assumption is false and Eq. (1) holds. ■

Lastly, we prove the second part of the theorem, Lemma 15, which states that any deterministic DFF algorithms has a mistake lower bound of at least DFFdim.

Proof [Proof of Lemma 15] Let $\mathcal{A}$ a deterministic DFF algorithm. Let $\texttt{Tr}$ be a shattered DFFT for $\mathcal{T}$ with history H, with height $d = \mathrm{DFFdim}(\mathcal{T}, H)$. We show that there exists a sequence of examples and a teacher for which $\mathcal{A}$ makes at least d mistakes, by traversing a path from the root of $\texttt{Tr}$ to a leaf and considering the interaction that it describes.

Denote the example at the root of the tree by x_1. Having observed x_1, the algorithm deterministically provides an explanation example $\hat{x}_1$ and a predicted label $\hat{y}_1$, where $(\hat{x}_1, \hat{y}_1) \in H$. Since $\texttt{Tr}$ is shattered by $\mathcal{T}, H$, then according to property 2 in Def. 3, the root has an outgoing edge labeled $(\hat{x}_1, \hat{y}_1)$. Let $v_2 = \langle y, \phi, x_2 \rangle$ be the target node of this edge. Set the teacher feedback for the algorihm's response to the label y with feature feedback ϕ. By property 1, $y \neq \hat{y}_1$. Thus, $\mathcal{A}$ makes a mistake in this round.

The next example in the sequence is set to x_2, which appears in v_2. This round and subsequent rounds proceed similarly: by the definition of a shattered DFFT, any pair of explanation and label provided by the algorithm is consistent with some outgoing edge of v_2. The teacher feedback and the next example are set by the target node v_3 of this edge, causing the algorithm to make another mistake. The run ends when reaching a leaf node. Since the height of $\texttt{Tr}$ is d and $\texttt{Tr}$ is complete (by property 4), this results in a sequence of d examples, on which $\mathcal{A}$ makes d mistakes. By property 3, the feedback provided in this run is consistent with some teacher in $\mathcal{T}$. This proves the claim. ■

Theorem 5 follows directly from Lemma 14 and Lemma 15.

Appendix B. Example: The DFFdim of a relaxed component model

The following DFF learning setting is a relaxed version of the component model first studied in Dasgupta et al. (2018). In our setting, each component is defined via a conjunction of features that hold for all examples in the component, as well as a common label. In addition, the label 0 is reserved for examples not in any component.

Example 2 (Relaxed component model) *Let $L, R, M \in \mathbb{N}$. Let $\mathcal{X}$ be an arbitrary set of examples, and let $\Phi \subseteq \{0,1\}^{\mathcal{X}}$ be an arbitrary set of features that is closed under negation. For a set $S \subseteq \Phi$, we say that x satisfies S if $\prod_{\phi \in S} \phi(x) = 1$ and denote $S(x) := \prod_{\phi \in S} \phi(x)$. Define $\mathcal{Y} = \{0, \ldots, L\}$. Each labeling function in the teacher class is defined by a collection of (at most) R sets of features $S_1, \ldots, S_R \subseteq \Phi$, where each set is of size (at most) M. Given $\mathcal{S} = \{S_1, \ldots, S_R\}$ and a mapping $q : [R] \to [L]$, the label of an example x is $q(j)$ if $S_j(x)$ holds. If more than one label satisfies this condition for any $x \in \mathcal{X}$, then the labeling function for $(\mathcal{S}, q)$ is undefined. If no j satisfies $S_j(x)$ then the label of x is set to 0. Denote the labeling function for $\mathcal{S}, q$ by $\ell_{\mathcal{S},q}$.*

Let $\mathcal{L} := \{(\mathcal{S}, q) \mid \ell_{\mathcal{S},q}$ is defined$\}$. For any $(\mathcal{S}, q) \in \mathcal{L}$, we define the set of possible teachers $\mathcal{T}_{\mathcal{S},q}$, by fixing their labeling function to $\ell_{\mathcal{S},q}$, but allowing different feature feedback functions ψ. The feature feedback function of each teacher in $\mathcal{T}_{\mathcal{S},q}$ satisfies the following: There exists a mapping $S : \mathcal{X} \to \mathcal{S}$ which selects the "primary" conjunction for each x with a non-zero label. Formally, for each x such that $\ell_{\mathcal{S},q}(x) = i \neq 0$, there is some j such that $S(x) = S_j$, $S_j(x)$ holds and $q(j) = i$. Denote $F(x) = \{\phi \mid \phi(x)$ holds.$\}$ For a set $A \subseteq \Phi$, denote $\neg A = \{\neg\phi \mid \phi \in A\}$. Then the ψ satisfies:

$$\text{If } \ell_{\mathcal{S},q}(\hat{x}) \neq 0, \psi(x, \hat{x}) \in F(x) \cap \neg S(\hat{x}). \text{ Otherwise, } \psi(x, \hat{x}) \in S(x) \cap \neg F(\hat{x}).$$

Note that this implies also that $\psi(x, \hat{x})$ satisfies x and does not satisfy $\hat{x}$, as required.

$\mathcal{T}_{\mathcal{S},q}$ is the set of all teachers with labeling function $\ell_{\mathcal{S},q}$ and a ψ function that satisfies the requirement above. The teacher class is set to $\mathcal{T} = \cup_{(\mathcal{S},q) \in \mathcal{L}} \mathcal{T}_{\mathcal{S},q}$.

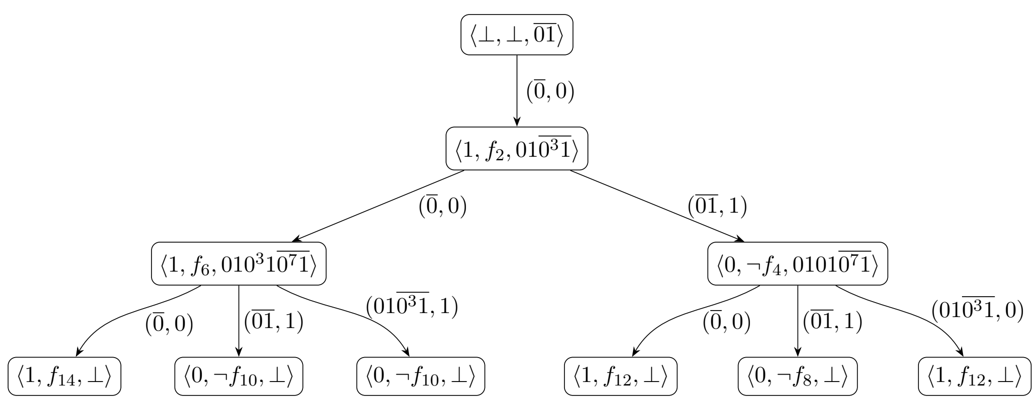

Figure 4: An example of a shattered DFFT of height 3 for the teacher class of the relaxed component model (Example 2) with $L = 1$, $R = 1$, $M = 3$, and $H = \{(\bar{0}, 0)\}$.

The following result provides an upper bound on the DFFdim of any example of the form in Example 2.

Theorem 17 *Given $L, R, M \in \mathbb{N}$, $\mathcal{X}$ and $\Phi \subseteq \{\texttt{true}, \texttt{false}\}^{\mathcal{X}}$, let $\mathcal{T}$ be defined as in Example 2. Let $H = \{(x_0, 0)\}$ for some $x_0 \in \mathcal{X}$ such that $\mathcal{T}$ is consistent with H. Then* $\mathrm{DFFdim}(\mathcal{T}, H) \leq RM$.

To prove this upper bound, we use a slight variant of the algorithm provided in Dasgupta et al. (2018) for the component model. The algorithm provided there assumes an initial arbitrary labeled example. The algorithm maintains a list of conjunctions attached to previously observed labeled examples. When a new example arrives, the algorithm finds a conjunction satisfied by the example and outputs the labeled example associated with this conjunction. If the predicted label is incorrect, the algorithm adds the negation of the resulting feature feedback to the conjunction of the labeled example it used for prediction. If no existing conjunction matches the presented example, the algorithm predicts using the default labeled example. If this prediction is incorrect, the algorithm creates a new conjunction that includes only the provided feature feedback and attaches to it the current example with its true label. The mistake bound of Dasgupta et al. (2018, Lemma 4) for this algorithm is proved for the (original) component model, and provides a bound of $R'M'$, where R' is the number of conjunctions created by the algorithm and M' is the maximal length of a conjunction created by the algorithm. The proof for the relaxed version in Example 2 is almost identical. The differences are minor, and are delineated in the proof sketch below.

Proof [Proof Sketch (Theorem 17)] The upper bound is obtained by showing the existence of a deterministic algorithm that obtains a mistake upper bound of RM for the given type of DFF problem. This is achieved by observing that the analysis that derives the mistake bound of the algorithm of Dasgupta et al. (2018) for the component model can be easily adapted for examples of type Example 2 as follows. The default labeled example assumed in Dasgupta et al. (2018) is replaced by the single labeled example $(x_0, 0) \in H$. Since a conjunction is never created for the default label by the algorithm, this guarantees that all conjunctions created by the algorithm are assigned an example that satisfies one of the conjunctions S_j and has a non-zero label. It follows from the definition of ψ for teachers in $\mathcal{T}$ that all features added to conjunctions by the algorithm do belong to the component $S(x)$ of the example x attached to the conjunction: if the conjunction is pre-existing, $x = \hat{x}$ and the feature is from $S(\hat{x})$, and if the conjunction is new then $x = x_t$ and the feature is from $S(x_t)$. We conclude that the maximal length M' of a conjunction created by the algorithm is M. In addition, the analysis of Dasgupta et al. (2018, Lemma 3) holds, showing that the number of rules R' is at most R. Thus, Dasgupta et al. (2018, Lemma 4) holds for Example 2 and provides a mistake bound of RM. ■

Using the definition of DFFdim, we can further show that for a given choice of $L, \mathcal{X}, \phi$ and H, the DFFdim for Example 2 is in fact equal to the worst-case RM, by constructing an appropriate shattered DFFT. This can be compared to the lower bound provided in Dasgupta and Sabato (2020), which is specific to the original component model and provides a lower bound equivalent to $RM/16$. That proof used a probabilistic analysis, which we can now avoid thanks to our new DFFdim construction.

Theorem 18 *Set $L = 1$, and set $\mathcal{X}, \Phi$ to be the natural feature construction defined in Section 5. For $\phi = f_n$ or $\phi = \neg f_n$, call n the coordinate of ϕ. Call f_n a positive feature and $\neg f_n$ a negative feature. Let $H = \{(\bar{0}, 0)\}$. Given $R, M \in \mathbb{N}$, define $S_1, \ldots, S_R \subseteq \Phi$ to be mutually exclusive sets of positive features f_n of size R. let $\mathcal{T}$ be as in Example 2. Then* $\mathrm{DFFdim}(\mathcal{T}, H) = RM$.

Proof The upper bound $\mathrm{DFFdim}(\mathcal{T}, H) \leq RM$ follows from Theorem 17. For the lower bound, we construct a shattered DFFT of height RM for $\mathcal{T}, H$. Set the root node of the DFFT to $\langle \perp, \perp, \overline{01} \rangle$. This node has a single child, with an edge labeled $(\bar{0}, 0)$.

We construct the rest of the tree inductively, such that in each path, each feature feedback is different from the previous ones in the path, and also consistent with all previous example labels. The sequence of features provided in each path from the root to a leaf determines $\mathcal{S}$ for a teacher that is consistent with this path.

Let A_v be the set of all $i \in \mathbb{N}$ such that for all labeled examples (x, y) in the path from the root to v, $x(i) = y$. Formally, the inductive construction maintains the following properties for every node $v = \langle \cdot, \cdot, x \rangle$.

- If $x \neq \perp$, then for any $y \in \{0, 1\}$, $|A_v \cap \{i \in \mathbb{N} \mid x(i) = y\}| = \infty$.
- Let $v_1, \ldots, v_t = v$ be the nodes in the path from the root to v, and denote $v_i = \langle y_{i-1}, \phi_{i-1}, x_i \rangle$. Let $I_v = (i_1, \ldots, i_{t-1})$ be the coordinates of the features $\phi_1, \ldots, \phi_{t-1}$. For any $i \in [t-1]$, let r_i be the maximal integer which is smaller than i and divisible by M. Then the following holds: $\forall j \leq r_i, x_i(i_j) = 0$, and $\forall j \in \{r_i + 1, \ldots, i-1\}, x_i(i_j) = 1$. In addition, for all $j \in \{i, \ldots, t-1\}$, $x_i(i_j) = y_i$. For instance, if $M = 3$ and $t = 10$ then $x_9(i_1, \ldots, i_{t-1}) = (0, 0, 0, 0, 0, 0, 1, 1, y_i, y_i)$.

These properties hold trivially for the root node. We inductively construct descendant nodes as follows. Let $v = \langle \cdot, \cdot, x \rangle$ be a node with outgoing edges and consider an outgoing edge from v labeled $(\hat{x}, \hat{y})$ where $\hat{y} \in \{0, 1\}$. Assume that the inductive hypothesis holds for v. We set the target node of this edge to $u = \langle y, \phi, x' \rangle$, where $y = 1 - \hat{y}$ and $\phi \in \Phi, x' \in \mathcal{X}$ are defined below.

Let $\bar{A}_u := A_u \setminus I_v$. These are feature coordinates that do not appear in the path from the root to v, whose value in all labeled examples on the path to u agrees with their labels. Note that $\bar{A}_u$ is infinite, since $A_u = A_v \cap \{i \in \mathbb{N} \mid x(i) = y\}$ is infinite by the induction hypothesis, and I_v is finite. Let $i = \min \bar{A}_u$. Then for any labeled example (x, y) in the path to u, $x(i) = y$. We set ϕ to f_i if $y = 1$ and to $\neg f_i$ if $y = 0$. Note that since $f_i(x) = x(i) = y$, in both cases $\phi(x) = 1$. In addition, since $(\hat{x}, \hat{y})$ is $(\bar{0}, 0)$ or a labeled example from the path to v, we have $f_i(\hat{x}) = \hat{x}(i) = \hat{y} = 1 - y$. Hence $\phi(\hat{x}) = 0$.

$x' \in \mathcal{X}$ is defined as follows. Let t be the depth of u. Let $I_u = (i_1, \ldots, i_t)$. If $t = RM$, we set $x' = \perp$ and u is set to be a leaf node. Otherwise, set $r \leq t$ to be the maximal integer that is divisible by M. Set $x'(i_j) = 0$ for all $j \leq r$, and $x'(i_j) = 1$ for all $j \in \{r+1, \ldots, t\}$. For coordinates in $\bar{A}_u \setminus I_u$, set their values to 0 and 1 in alternating order, so that infinitely many of such coordinates have a value of 0 and infinitely many of them have a value of 1. Other coordinates of x' can have arbitrary values. Define an outgoing edge from u for every pair $(\hat{x}, \hat{y})$ in the labeled examples to u or in H. An illustration of some of the construction of x' is provided in Figure 5.

It can now be verified that the induction hypothesis holds for this construction: $A_u \cap \{i \in \mathbb{N} \mid x'(i) = y\}$ is infinite from the choice of x'. The second property holds for x' from the choice of x', and for other examples on the path from the inductive hypothesis and the choice of ϕ as agreeing with all previous labeled examples.

To prove that the result is a shattered DFFT for $\mathcal{T}$ and H, it suffices to show that for any path from the root to a leaf, there is a teacher in $\mathcal{T}$ that is consistent with this path. Let v be the leaf. Partition I_v into R sub-sequences of size M, and define $\mathcal{S} = \{S_1, \ldots, S_R\}$, where S_i includes the positive features of the coordinates in the i'th set of the partition. Set q to the constant 1 function.

$$
\begin{array}{llll}
(\bot,\bot,x_1=\overline{01}) & \xrightarrow{(\bar{0},0)} & (1, f_2, & x_2 = \star 1 \overline{\star 0 \star 1}) \\
& \xrightarrow{(x_1,1)} & (0, \neg f_4, & x_3 = \star 1 \star 1 \overline{\star^3 0 \star^3 1}) \\
& \xrightarrow{(x_2,0)} & (1, f_{12}, & x_4 = \star 0 \star 0 \star^7 0 \overline{\star^7 0 \star^7 1}) \\
& \xrightarrow{(x_3,1)} & (0, \neg f_{28}, & x_5 = \star 0 \star 0 \star^7 0 \star^{15} 1 \overline{\star^{15} 0 \star^{15} 1}) \ldots
\end{array}
$$

Figure 5: An example of a path prefix starting from the root in the inductive construction of the tree in the proof of Theorem 18, for $M = 3$. The symbol '$\star$' indicates an arbitrary Boolean value. An overline indicates infinite repetition of the sequence under it.

Then from the second property of the induction hypothesis, it is easy to verify that $\ell_{\mathcal{S},q}(x) = y$ for all labeled examples (x, y) in the path. Moreover, the feature feedback provided along the path is consistent with some teacher in $\mathcal{T}_{\mathcal{S},q}$, since it satisfies the properties required from ψ in Example 2. We have thus constructed a shattered DFFT of height RM for $\mathcal{T}, H$. ■

Appendix C. Proofs of Theorems 8, 9: dimensions in online conversion

Proof [of Theorem 8] Denote $(\mathcal{T}_\mathcal{F}, H) = \mathrm{OtD}(\mathcal{F})$. Then $\mathcal{X} \cap H_\mathcal{X} = \emptyset$, and $\mathcal{T}_\mathcal{F}$ is a teacher class over $\mathcal{X} \cup H_\mathcal{X}, \mathcal{Y}$ and some Φ. Denote $\mathcal{F}' = \mathrm{DtO}(\mathcal{T}_\mathcal{F}, H)$. Then $\mathcal{F}'$ is a hypothesis class over $(\mathcal{X} \cup H_\mathcal{X}) \setminus H_\mathcal{X} = \mathcal{X}$. Therefore, $\mathcal{F}$ and $\mathcal{F}'$ are both subsets of $\mathcal{Y}^\mathcal{X}$.

To show that $\mathcal{F} = \mathcal{F}'$, let $f \in \mathcal{F}$. By the definition of OtD, there is a teacher $T = (\ell, \psi) \in \mathcal{T}_\mathcal{F}$ such that $\ell|_\mathcal{X} = f$. By the definition of DtO, it follows that there is some $f' \in \mathcal{F}'$ such that $f' = \ell|_\mathcal{X} = f$. It follows that $\mathcal{F} \subseteq \mathcal{F}'$. Conversely, for any $f' \in \mathcal{F}'$, there is some $T = (\ell, \psi) \in \mathcal{T}_\mathcal{F}$ such that $\ell|_\mathcal{X} = f'$. It follows from the definition of OtD that there exists some $f \in \mathcal{F}$ such that $\ell|_\mathcal{X} = f$. Therefore, $\mathcal{F}' \subseteq \mathcal{F}$. It follows that $\mathcal{F} = \mathcal{F}'$. ■

To prove Theorem 9, we first provide two lemmas. The first lemma shows that the mistake bound cannot decrease by the conversion. The second lemma provides a lower bound for a mapping in the converse direction.

Lemma 19 *For any hypothesis class $\mathcal{F} \in \mathcal{Y}^\mathcal{X}$,* $\mathrm{DFFdim}(\mathrm{OtD}(\mathcal{F})) \geq \mathrm{Ldim}(\mathcal{F})$.

Proof Denote $(\mathcal{T}_\mathcal{F}, H) := \mathrm{OtD}(\mathcal{F})$. To prove the claim, we construct a DFFT for $\mathcal{T}_\mathcal{F}$ and H of height at least $\mathrm{Ldim}(\mathcal{F})$.

Let $\texttt{Tr}$ be a shattered Littlestone tree for $\mathcal{F}$ with height $\mathrm{Ldim}(\mathcal{F})$. This is a full binary tree with internal nodes labeled by examples and We construct $\texttt{Tr}'$, a shattered DFFT of the same height, by inductively mapping from nodes in $\texttt{Tr}$ to nodes in $\texttt{Tr}'$. For a node v in $\texttt{Tr}'$, let $N(v)$, which will be defined inductively, be the node in $\texttt{Tr}$ that it was mapped from.

Assume that $\texttt{Tr}$ is of height at least 1, otherwise the statement holds trivially. The root node v_0' of $\texttt{Tr}'$ is set to $\langle \bot, \bot, x_1 \rangle$, where x_1 is the example in the root node v_0 of $\texttt{Tr}$. We set $N(v_0') := v_0$. Inductively from the root, for every node $v = \langle \cdot, \cdot, x \rangle$ in $\texttt{Tr}'$ such that $x \neq \bot$, define outgoing edges labeled by $(\hat{x}, \hat{y}) \in \mathcal{X} \times \mathcal{Y}$, for each possible pair as defined in Property 2 of Def. 3. Note that by definition, there is always at least one such pair. Let y_1, y_2 be the two outgoing edges from $N(v)$

in Tr, and let x_1, x_2 be the respective examples in the target nodes u_1, u_2 of these edges in Tr (see Figure 2). Note that if u_1, u_2 are leaves in Tr then they are not labeled by an example. In this case, set x_1, x_2 to $\perp$. In Tr$'$, set the target of each outgoing edge $(\hat{x}, \hat{y})$ to a node $\langle y_1, \mathbb{I}[x], x_1\rangle$ if $\hat{y} \neq y_1$, and to $\langle y_2, \mathbb{I}[x], x_2\rangle$ if $\hat{y} = y_1$. For each such node v', we define $N(v') = u_i$ for the appropriate $i \in \{1, 2\}$. Note that more than one node in Tr$'$ can be mapped from the same node in Tr.

It is easy to verify that the resulting tree Tr$'$ is a DFFT, and that it satisfies the properties required from a shattered tree. In particular, for every path in Tr$'$, there is a teacher in $\mathcal{T}$ that is consistent with it: The sequence of labeled examples in each of the paths exists also as a path in Tr, therefore there exists a function $f \in \mathcal{F}$ that is consistent with this sequence. The labeling function provided by the teacher $T_f \in \mathcal{T}$ is thus consistent with the labels on the path in Tr$'$. In addition, by definition of T_f and Tr$'$, the features in each node in the path are also consistent with T_f. Lastly, all teachers in $\mathcal{T}$ are consistent with H, as required. Since Tr$'$ is a shattered DFFT with the same height as Tr, we have $\mathrm{DFFdim}(\mathrm{OtD}(\mathcal{F})) \geq \mathrm{Ldim}(\mathcal{F})$. ■

Next, we show lower-bound on the Littlestone dimension for the converse conversion.

Lemma 20 *Given a teacher class $\mathcal{T}$ and a history H, let N be the number of labels from $\mathcal{Y}$ that do not appear in any pair in H. Then*

$$\mathrm{Ldim}(\mathrm{DtO}(\mathcal{T}, H)) \geq \mathrm{DFFdim}(\mathcal{T}, H) - N.$$

Proof Denote $\mathcal{F} := \mathrm{DtO}(\mathcal{T}, H)$. To prove the claim, we construct a shattered Littlestone tree for $\mathcal{F}$ of height $\mathrm{DFFdim}(\mathcal{T}, H) + |\mathcal{Y}| - \{y \mid (\cdot, y) \in H\}$.

Let Tr be a shattered DFFT of height $\mathrm{DFFdim}(\mathcal{T}, H)$. Assume that Tr is of height at least 1, otherwise the statement holds trivially. We construct a Littlestone tree Tr$'$ for $\mathcal{F}$ inductively, as follows. For a node v in Tr$'$, let $N(v)$, which will be defined inductively, be the node in Tr that it was mapped from.

Label the root node u of Tr$'$ with the example in the root v of Tr. We set $N(u) = v$. Inductively from the root, for a node u in Tr$'$ which was mapped from some v, consider the children of v in Tr. If there are two child nodes of v with two different labels $y \neq y'$, then arbitrarily select two such nodes $\langle y, \phi, x\rangle$ and $\langle y', \phi', x'\rangle$ and set two outgoing edges from u labeled y and y'. Label the target nodes of these edges by x and x', respectively. If $x, x' = \perp$, then set the target nodes of the edges to unlabeled leaf nodes.

If all the child nodes of v are labeled with the same label, arbitrarily select one of v's child nodes, and use its child nodes instead of v's child nodes. This may be repeated, until a descendant of v with two child nodes with different labels is found. After constructing the whole tree, prune paths from root to leaf in Tr$'$ so that they all have the length of the minimal-length path, resulting in a complete binary tree.

To verify that Tr$'$ is a Littlestone tree, note that each path in Tr corresponds to the labeling function of some teacher in $\mathcal{T}$, and therefore, by the definition of DtO, also to some function in $\mathcal{F}$. Moreover, the height of Tr$'$ is at least $\mathrm{DFFdim}(\mathcal{T}, H) - N$. This is because the only time that a path in Tr is shortened by 1 for Tr$'$ is if all labels of a node v's child nodes are the same label y. However, by the definition of a shattered DFFT, this can only be the case if none of the outgoing edges from v are labeled with y. This is only possible if y is not a label in H and also it is not in any labeled example in the path to v. However, after one occurrence of the latter, y is now in a labeled example in the path. Therefore, each label not in H can cause at most a decrease of 1 in the length of the path. This completes the proof. ■

Using the lemmas above, Theorem 9 can be easily proved.

Proof [Proof of Theorem 9] First, by Lemma 19, $\mathrm{Ldim}(\mathcal{F}) \le \mathrm{DFFdim}(\mathrm{OtD}(\mathcal{F}))$. For the other direction of the inequality, Let $(\mathcal{T}, H) = \mathrm{OtD}(\mathcal{F})$. By the definition of OtD, all the labels in $\mathcal{Y}$ appear in H. Therefore, by Lemma 20, $\mathrm{Ldim}(\mathrm{DtO}(\mathcal{T}, H)) \ge \mathrm{DFFdim}(\mathcal{T}, H)$. Equivalently, $\mathrm{Ldim}(\mathrm{DtO}(\mathrm{OtD}(\mathcal{F})) \ge \mathrm{DFFdim}(\mathrm{OtD}(\mathcal{F}))$. By Theorem 8, $\mathrm{DtO}(\mathrm{OtD}(\mathcal{F})) = \mathcal{F}$, hence $\mathrm{Ldim}(\mathcal{F}) \ge \mathrm{DFFdim}(\mathrm{OtD}(\mathcal{F}))$. ■

Appendix D. A non-realizable mistake upper bound for general interactive protocols

We state and prove Theorem 21, stated informally above as Theorem 12. Theorem 21 is a generalization of the k-non-realizable upper bound of Theorem 11 to a wide range of interactive protocols. We consider a general interactive prediction protocol in which each round t is of the following form:

1. The environment provides some x_t;
2. The algorithm outputs z_t;
3. The environment provides feedback w_t.

We further assume some initial input I that is provided before the start of the first round, and a class of generalized teachers $\mathcal{T}$ (not to be confused with the specific notion of teachers in DFF), whose role is defined below. The protocol defines the following rules of interaction between the teacher and the algorithm.

- The legal responses of the algorithm, given the example presented in the current round and the run so far. Formally, given $D = \{I, x_t, ((x_1, z_1, w_1), \ldots, (x_{t-1}, z_{t-1}, w_{t-1}))\}$, the protocol defines $A(D)$, the set of values that the algorithm may provide as z_t.
- For every possible teacher T, the protocol defines for each x_t and z_t which feedback values w_t are consistent with T.

Many interactive protocols fall under this definition, including, for instance, Online Learning, Online Selective Classification with Limited Feedback (Gangrade et al., 2021), bandit Online Learning (Daniely et al., 2015), and learning with per-sample side information (Visotsky et al., 2019). In particular, DFF is an instance of this general protocol, with $I := H$.

Theorem 21 is a generalized form of Theorem 11, which holds for interactive protocols as defined above that satisfy the following natural assumptions. These assumptions hold for DFF, as well as the other protocols mentioned above.

Assumption D.1 *There exists a fixed known function that given (x_t, z_t, w_t) for a given round t, indicates whether the algorithm made a mistake in this round.*

Assumption D.2 *The outputs z_t that the algorithm is allowed to provide if executed on a suffix of another run, are a subset of those allowed in the original run. Formally, let $t, l \in \mathbb{N}$ such that $t \ge l$. Given I, $\{(x_i, z_i, w_i)\}_{i \in [t-1]}$ and x_t, let*

$$D = \{I, x_t, ((x_1, z_1, w_1), \ldots, (x_{t-1}, z_{t-1}, w_{t-1}))\},$$
$$D' = \{I, x_t, ((x_l, z_l, w_l), \ldots, (x_{t-1}, z_{t-1}, w_{t-1}))\}.$$

Then $A(D') \subseteq A(D)$.

In a run with L rounds, the number of protocol exceptions is the size of the smallest subset $E \subseteq [L]$ such that the feedbacks w_t for $t \in [L] \setminus E$ are consistent with some $T \in \mathcal{T}$. Now, suppose that there is an algorithm $\mathcal{A}$ for the given protocol and a function $\mathcal{M}$ that maps teacher classes to natural numbers, such that $\mathcal{A}(\mathcal{T}, I)$ makes at most $\mathcal{M}(\mathcal{T}, I)$ incorrect predictions in a run with zero exceptions. In the case of DFF, $\mathcal{A}$ can be set to `SOA-DFF` and $\mathcal{M}(\mathcal{T}, I)$ to $\mathrm{DFFdim}(\mathcal{T}, I)$. To handle exceptions, one can use the algorithm in `GAA` listed in Alg. 3, which is a general form of `A-SOA-DFF` presented in Section 6.

Algorithm 3 Generic Agnostic Algorithm

procedure `GAA`($\mathcal{A}, \mathcal{T}, I$)
    Initialize $\mathcal{A}(\mathcal{T}, I)$ and set $M \leftarrow 0$.
    for $t = 1, 2, \ldots$ **do**
        Receive x_t and provide it to $\mathcal{A}$; Record the z_t provided by $\mathcal{A}$.
        Output z_t
        Receive feedback w_t.
        If (x_t, z_t, w_t) indicate a mistake, $M \leftarrow M + 1$.
        If $M = \mathcal{M}(\mathcal{T}, H) + 1$, re-initialize $\mathcal{A}(\mathcal{T}, I)$ and set $M \leftarrow 0$

Theorem 21 *If the interactive protocol satisfies Assumptions D.1 and D.2, then in a run with at most k exceptions,* `GAA` *makes at most $(k+1)\mathcal{M}(\mathcal{T}, I) + k$ mistakes.*

Proof First, observe that by D.1 $\mathcal{A}$ can use (x_t, z_t, w_t) to identify mistakes. Moreover, if $\mathcal{A}$ is a valid algorithm for the given interactive protocol, then under Assumption D.2 so is `GAA`($\mathcal{A}, \mathcal{T}, I$).

Call each sub-sequence of the run between initializations of $\mathcal{A}$ a "run segment" and denote $d = M(\mathcal{T}, I)$. In each run segment, $\mathcal{A}$ makes $d + 1$ mistakes. Let N be the total number of run segments. Then the total number of mistakes made by the algorithm is at most $N(d+1) + d$. Since the realizable mistake bound of $\mathcal{A}(P)$ is d, all teachers in $\mathcal{T}$ are inconsistent with all of the N run segments. Therefore, any such teacher is inconsistent with at least N rounds in the whole run. If there are at most k exceptions during the run, it follows that $N \leq k$. Thus, `GAA`($\mathcal{A}, \mathcal{T}, I$) makes at most $k(d+1) + d = (k+1)d + k$ mistakes. ■

Appendix E. Proof of Theorem 13: a non-realizable lower bound

In this section, we provide the full proof of Theorem 13. In Section E.1 we give background on the secret-sharing scheme that we are using. We describe the construction of the teacher classes in Section E.2. We prove an upper bound on the DFFdim of these classes in Section E.3, Lemma 22. In Section E.4, we prove Lemma 23, which provides a lower bound on the mistake bound of any algorithm on these classes. Theorem 13 directly follows from Lemma 22 and Lemma 23.

E.1. Background: The secret-sharing scheme

The secret sharing scheme of Shamir (1979) works as follows: Let $\mathbb{F}_p$ denote the field of size p for some prime number p, and consider a polynomial over $\mathbb{F}_p$ of degree $d-1$ with coefficients in $\mathbb{F}_p$, $P := (c_0 + \sum_{j \in [d-1]} c_j x^j) \bmod p$. Given a secret $s \in \mathbb{F}_p$, arbitrary coefficients $c_1, \ldots, c_{d-1} \in \mathbb{F}_p$

are selected, and we define $c_0 = s$, so that $P(0) = s$. The partial secrets are then defined as $\bar{s} = \{s_i\}_{i \in [p]}$, where $s_i = (i, P(i))$.

According to Lagrange's interpolation theorem, the polynomial P can be reconstructed from any d shares of the secret $s_{i_j} = (i_1, z_{i_j})$ for $j \in [d]$ via the Lagrange interpolating polynomial:

$$P(x) = \sum_{j=1}^{d} z_{i_j} \prod_{l \in [d]: l \neq j} \frac{x - i_l}{i_j - i_l} \bmod p.$$

Lagrange's interpolation theorem states that this polynomial is the unique polynomial of degree $d-1$ that satisfies $\forall j \in [d]$: $P(i_j) = z_{i_j}$. Once P is found, one can obtain the secret s by evaluating $P(0)$. However, for any set of $d-1$ partial secrets $s_{i'_1}, \ldots, s_{i'_{d-1}}$, and every $s' \in \mathbb{F}_p$, there exists a polynomial P' with $P'(0) = s'$ such that for all $j \in [d-1]$, $P'(i'_j) = P(i'_j)$. Thus, informally, it is not possible to infer anything useful on s from $d-1$ partial secrets. This mapping thus describes a (d, p)-threshold secret sharing scheme.

E.2. Construction of the teacher class

Given a natural number $d \geq 1$, we define the teacher class $\mathcal{T}^{d+1}$ as follows. Let the domain $\mathcal{X} = \mathbb{N}$ and the label set be $\mathcal{Y} = \{0, 1\}$. Set the feature set to $\Phi = \{0, 1\}^{\mathcal{X}}$. Partition the set $\mathcal{X}$ into mutually exclusive sets $\mathcal{X}_i = [2^{i+1} - 1] \setminus [2^i - 1]$, such that $\mathcal{X} = \bigcup_{i=0}^{\infty} \mathcal{X}_i$. We construct a teacher class that employs the secret-sharing idea described in Section 6 above. We define $\mathcal{T}^{d+1}$ by first defining teacher classes $\mathcal{T}_i^d$ for $i \in \mathbb{N}$. For a given i, we define a class of labeling functions $\mathcal{F}_i^d$. Then, for a given $f \in \mathcal{F}_i^d$, we define the set of all compatible feature feedback functions Ψ_f. We then define $\mathcal{T}_i^d = \{(f, \psi) : f \in \mathcal{F}_i^d, \psi \in \Psi_f\}$, and set $\mathcal{T}^{d+1} = \cup_{i=1}^{\infty} \mathcal{T}_i^d$.

The class $\mathcal{F}_i^d$ includes all the functions $f : \mathcal{X} \to \{0, 1\}$ that satisfy the following:

- $f(1) = 1$
- For all $x \in \mathcal{X} \setminus (\mathcal{X}_i \cup \{1\})$, $f(x) = 0$.

Thus, there is exactly one function in $\mathcal{F}_i^d$ for each possible labeling of $\mathcal{X}_i$.

For a given f, we construct Ψ_f, the class of possible feature-feedback functions, such that the Ψ_f are mutually exclusive over all $f \in \cup_{i=0}^{\infty} \mathcal{F}_i^d$ for any i. To satisfy the construction properties as described in Section 6, Let p_i be a prime number such that $p_i \geq |\mathcal{F}_i^d| = 2^{|\mathcal{X}_i|}$. We construct Ψ_f to correspond to a (d, p_i)-threshold secret sharing scheme for $f \in \mathbb{F}_i$.

We first present the general form of the functions in Ψ_f, using secret-sharing polynomials. For coefficients $\bar{c} = (c_0, \ldots, c_{d-1}) \in \mathbb{F}_{p_i}^d$, define the secret-sharing polynomial of degree $d-1$ over $\mathbb{F}_{p_i}$, $P'[\bar{c}] := (c_0 + \sum_{j \in [d-1]} c_j x^j) \bmod p_i$, whose secret is $P'(0) = c_0$. We then define a "shifted" version of this polynomial, $P[\bar{c}] : \mathcal{X}_i \to \mathcal{X} \setminus \mathcal{X}_i$, via

$$P[\bar{c}](x) = 2^{i+1} + P'[\bar{c}](x - 2^i + 1).$$

For given elements $x_1, \ldots, x_l \in \mathcal{X}$, denote by $\mathbb{I}[x_1, \ldots, x_l] : \mathcal{X} \to \{0, 1\}$ the feature that holds if and only if $x \in \{x_1, \ldots, x_l\}$. Now, given a function $\bar{a} : \mathcal{X}_i \times \{0, 1\} \to \mathbb{F}_{p_i}^d$, define the feature feedback function $\psi_{\bar{a}}^f : \mathcal{X} \times \mathcal{X} \to \Phi$ by

$$\psi_{\bar{a}}^f(x, \hat{x}) = \begin{cases} \mathbb{I}[x, P[\bar{a}(\hat{x}, f(x))](x)] & x, \hat{x} \in \mathcal{X}_i \\ \mathbb{I}[x] & \text{otherwise.} \end{cases} \quad (2)$$

Thus, the feature feedback function defines a separate secret polynomial for every $\hat{x}$ and every label of x, and provides an evaluation of this polynomial on x. Note that the first and second element in the first case are never the same, since the shifted polynomials' outputs are not in $\mathcal{X}_i$. This also implies that if $x \neq \hat{x}$ then $\psi^f_{\bar{a}}(x, \hat{x})$ satisfies x and does not satisfy $\hat{x}$, as required.

To define the set Ψ_f, we encode the hypotheses in $\mathcal{F}^d_i$ as numbers via a fixed bijective function $b_i : \mathcal{F}^d_i \to \mathbb{F}_{p_i}$. We then include in Ψ_f feature feedback functions that encode $b_i(f)$ in some of its secret-sharing polynomials. The rest of the polynomials encode an arbitrary "fake" secret.

$$\Psi_f = \{\psi^f_{\bar{a}} \mid \bar{a} : \mathcal{X}_i \times \{0, 1\} \to \mathbb{F}^d_{p_i}, \text{ and } \forall \hat{x} \in \mathcal{X}_i, f(\hat{x}) = 1 \Rightarrow \bar{a}(\hat{x}, 0)(0) = b_i(f).\}.$$

Thus, for feature feedback functions in Ψ_f, a part of the secret $b_i(f)$ is revealed given $x, \hat{x}$ if and only if $f(\hat{x}) = 1$ and $f(x) = 0$. The rest of the cases reveal evaluations of a "fake" polynomial that does not encode the secret. We show below that this construction guarantees that the algorithm must observe d true feature feedback responses for the same explanation $\hat{x}$ to reveal any information on f beyond the labels of observed examples.

E.3. Upper bounding the DFFdim of the construction

The following result gives an upper bound on the DFF dimension of the teacher classes defined in the construction above. Recall that $H_\bullet = \{(2, 0), (1, 1)\}$, as defined in Theorem 13.

Lemma 22 *Let $i \in \mathbb{N}$. Suppose that $H = \{(x^0_0, 0), (x^1_0, 1)\}$ for some $x^0_0, x^1_0 \in \mathbb{N}$ such that H is consistent with some teacher in $\mathcal{T}_i \subseteq \mathcal{T}^{d+1}$ for some $i \in \mathbb{N}$. Then $\mathrm{DFFdim}(\mathcal{T}^d_i, H) \leq d$ and $\mathrm{DFFdim}(\mathcal{T}^{d+1}, H) \leq d + 1$. In particular, the history $H_\bullet$ satisfies this for all $i \in \mathbb{N}$.*

Proof We prove the result for $\mathcal{T}^{d+1}$ by providing a deterministic algorithm for this class with a mistake bound of $d + 1$ in the realizable case. By Theorem 5, this implies that $\mathrm{DFFdim}(\mathcal{T}^{d+1}) \leq d + 1$. We then show how to adapt the algorithm to obtain a mistake bound of d if the teacher class is $\mathcal{T}^d_i$.

The algorithm for $\mathcal{T}^{d+1}$, denoted $\mathcal{A}$, is provided in Alg. 4. The algorithm predicts 0 for any unknown example until the first example with a label of 1 that is not $x = 1$ is observed (note that if $x^1_0 \neq 1$, this occurs as soon as the algorithm starts). Denote the true teacher by $T^* = (f^*, \psi^{f^*}_{\bar{a}})$. Once n is defined, for any example with an unknown label, the algorithm predicts 1, using this first example as explanation. It collects the partial secret obtained from the feature feedback for each mistake, and after collecting d such parts, it reconstructs the true labeling function and uses it to label all subsequent examples. We write "output (x,y)" to indicate that the algorithm predicts label y with explanation x.

We now prove that this algorithm makes at most $d+1$ mistakes. First, the algorithm predicts 0 for all previously unobserved examples (except for $x_t = 1$) until making its first mistake. Therefore, this mistake reveals an example $x_t \neq 1$ such that $f^*(x_t) = 1$. In this round, the algorithm sets $\hat{x} \leftarrow x_t$ and n is set such that $\hat{x} \in \mathcal{X}_n$ (line 9). It follows that $f^* \in \mathcal{F}_n$. Therefore, after this round, the only unknown labels are of examples in $\mathcal{X}_n$, and the algorithm makes mistakes only on such examples.

Now, suppose that $|R| < d$ and a new example x_t from $\mathcal{X}_n$ is observed. The algorithm outputs $(\hat{x}, 1)$. If this is a mistake, then $f^*(x_t) = 0$ and the algorithm received a feature feedback $\phi_t = \psi^{f^*}_{\bar{a},\hat{x}}(\hat{x}, x_t) = \mathbb{I}[x, P[\bar{a}(\hat{x}, 0)](x_t)]$ that provides an evaluation of $P[\bar{a}(\hat{x}, 0)]$ for an x_t that has

Algorithm 4 A Learning algorithm for $\mathcal{T}^{d+1}$, $H = \{(x_0^0, 0), (x_0^1, 1)\}$.

Initialize $R \leftarrow \emptyset$. If $x_0^1 \neq 1$, set $\hat{x} \leftarrow x_0^1$ and define n to the unique integer such that $\hat{x} \in \mathcal{X}_n$.
for $t = 1, ..., L$ **do**
  if $\hat{f}$ is defined **then** output $(x_0^{\hat{f}(x_t)}, \hat{f}(x_t))$.
  else if $x_t = 1$ **then** output $(x_0^1, 1)$.
  else if there is a positive integer $s < t$ such that $x_s = x_t$ **then** output (x_s, y_s).
  else if $|R| < d$ **then**
    if n is undefined **then**
      output $(x_0^0, 0)$.
      if mistake **then** set $\hat{x} \leftarrow x_t$ and define n to the unique integer with $x_t \in \mathcal{X}_n$.
    else if $x_t \in \mathcal{X}_n$ **then**
      output $(\hat{x}, 1)$.
      if mistake **then**
        receive feature feedback $\mathbb{I}[x_t, z_t]$,
        update $R \leftarrow R \cup \{(x_t - 2^n + 1, z_t - 2^{n+1})\}$.
    else (n is defined, $x_t \notin \mathcal{X}_n$)
      output $(x_0^0, 0)$.
      If mistake, ignore feature feedback.
  else ($|R| = d$)
    Denote $R = \{(\tilde{x}_1, \tilde{z}_1), \ldots, (\tilde{x}_d, \tilde{z}_d)\}$.
    Define the polynomial $\hat{P}$ via Lagrange interpolation:

$$\hat{P}(x) := \sum_{j=1}^{d} \tilde{z}_j \prod_{l \in [d]: l \neq j} \frac{x - \tilde{x}_j}{\tilde{x}_l - \tilde{x}_j} \bmod p_n.$$

    Define $\hat{f} \leftarrow b_n^{-1}(\hat{P}(0))$.
    Output $(x_0^{\hat{f}(x_t)}, \hat{f}(x_t))$.

never been seen before. At most d such rounds are possible. If $|R| = d$, then the algorithm uses the d feedbacks received so far to reconstruct f^*. This can be done because R includes d distinct evaluations of the same polynomial $P[\bar{a}, (\hat{x}, 0)]$, where $\bar{a}(\hat{x}, 0) = b_n(f^*)$, since $f^*(\hat{x}) = 1$. Thereafter, the algorithm makes no more mistakes, since $\hat{f} = b^{-1}(\hat{P}(0)) = f^*$ for $\hat{P} = P'[\bar{a}, (\hat{x}, 0)]$. Taking everything together, we see that the algorithm can make at most $d + 1$ mistakes.

To complete the proof, the above algorithm can be adapted for the teacher class $\mathcal{T}_i$ by initializing n to i. In this case, the algorithm makes at most d mistakes since it does not make the mistake in the round where n is set. ∎

E.4. Lower-bounding the number of mistakes of the construction

In this section, we prove a worst-case lower bound for the number of mistakes any algorithm would make when learning the class $\mathcal{T}^{d+1}$ defined in Section E.2, with the history $H_\bullet = \{(2, 0), (1, 1)\}$ defined in Theorem 13. This lower bound holds for adaptive adversaries. To make this precise, we introduce our assumptions on the interaction between the algorithm and the adversary below. We denote the sequence of interactions over the first t rounds of a run by $S_t = ((x_1, y_1, \hat{x}_1, \hat{y}_1, \phi_1), \ldots, (x_{t-1}, y_{t-1}, \hat{x}_t, \hat{y}_t, \phi_t))$, where we denote $\phi_i = \perp$ whenever $\hat{y}_i = y_i$. We assume the following order of decisions by the algorithm and the adversary: In each round t, the adversary first selects (x_t, y_t, ψ_t), where $\psi_t : \mathcal{X} \times \mathcal{X} \to \Phi$. The choice of this triplet can be dependent on S_{t-1}. Then the algorithm selects $(\hat{x}_t, \hat{y}_t)$, where its decision can depend on S_{t-1}, x_t. Lastly, $\phi_t = \psi_t(x_t, \hat{x}_t)$ is presented to the algorithm. The algorithm can be randomized, thus $(\hat{x}_t, \hat{y}_t)$ can be drawn from a distribution that depends on S_{t-1}, x_t.

Let L be an integer. For a sequence $S := S_L$, S is said to be k-consistent with a teacher $T = (f, \psi)$ if there exists a set $E \subseteq [L]$ such that $|E| \leq k$ and for all $t \in [L] \setminus E$, we have $f(x_t) = y_t$, and if $\hat{y}_t \neq y_t$ then $\psi(x_t, \hat{x}_t) = \phi_t$. For a given algorithm $\mathcal{A}$ and an adversary V, let $M^L(\mathcal{A}, V, H)$ be the expected number of prediction mistakes made by $\mathcal{A}$ when interacting with V over a sequence of length L. The expected worst-case number of mistakes of a randomized algorithm $\mathcal{A}$ in the k-non-realizable setting for a teacher class $\mathcal{T}$ is defined as

$$M_k^L(\mathcal{A}, \mathcal{T}, H) = \sup_{V \in \mathcal{V}(\mathcal{T}, \mathcal{A})} M^L(\mathcal{A}, V, H),$$

where $\mathcal{V}(\mathcal{T}, \mathcal{A})$ is the set of adversaries such that when they interact with $\mathcal{A}$, with probability 1 the generated sequence S is k-consistent with some teacher $T \in \mathcal{T}$.

To compare this model to the randomized model of Online Learning, note that DFF with uninformative explanations is equivalent to Online Learning, as shown in Section 5. When applying the DFF adversary defined above to Online Learning via this mapping, this recovers Online Learning with an *adaptive* adversary (as defined in Arora et al., 2012) who makes up to k labeling errors. An adaptive adversary for Online Learning obtains the same optimal mistake bound of $k + \Theta\left(\sqrt{k \cdot \mathrm{Ldim}(\mathcal{F})} + \mathrm{Ldim}(\mathcal{F})\right)$ as an adversary who fixes the sequence in advance, using the Weighted Majority algorithm (Littlestone and Warmuth, 1994). Thus, our lower bound should be compared to this lower and upper bound for Online Learning.

We prove the following lower bound for the k-non-realizable setting.

Lemma 23 *Let $k, d \geq 1$ and let $L \geq 4(k+1)d$. Let $\mathcal{A}$ be a (possibly randomized) algorithm. Then*

$$M_k^L(\mathcal{A}, \mathcal{T}^{d+1}, H_\bullet) \geq (k+1)d - 1.$$

We prove this result by presenting a strategy for the adversary V given a learning algorithm $\mathcal{A}$, as well as L and k. We then show that this adversary only generates sequences that are k-consistent with $\mathcal{T}^{d+1}$, and prove a lower bound for $M_k^L(\mathcal{A}, V, H_\bullet)$.

Given a sequence S, define the version space $\mathcal{T}_S = \{T \in \mathcal{T} \mid S \text{ is (fully) consistent with } T\}$. For a given set of exception rounds $E \subseteq [L]$, denote $S^{\setminus E} = ((x_t, y_t, \hat{x}_t, \hat{y}_t, \phi_t))_{t \in [L] \setminus E}$. We further denote $S_{\mathcal{X}} := \{x_i\}_{i \in [L]}$. Lastly, for a sequence S we denote the set of rounds in which the same explanation $\hat{x}$ is given for examples with the same true label y, and in which a feature feedback is provided, by

$$R(S, \hat{x}, y) := \{t \in [L] \mid \hat{x}_t = \hat{x}, y_t = y \text{ and } \phi_t \neq \perp\}.$$

The adversary V. Given $\mathcal{A}$, k and $L \geq 2$, the adversary V is defined as follows. Fix $i := L$. V only presents examples x_t from $\mathcal{X}_i$, and the generated sequences will all be k-consistent with teachers in $\mathcal{T} := \mathcal{T}_i^d$. More precisely, V will be k-consistent with a teacher for which the exception rounds are $E_L \subseteq [L]$, where E_t for $t \in [L]$ is defined inductively throughout the run. Set $E_0 := \emptyset$. E_t for $t \geq 1$ is defined inductively below. In round t, the adversary makes the following choices:

- $x_t := 2^i + t$.
- We now define the choice of y_t and ψ_t. For $y \in \{0, 1\}$, denote the teachers consistent with rounds not declared as exceptions so far, as well as with y as a label for x_t, by

  $$\mathcal{T}_y = \{T = (f, \psi) \in \mathcal{T} \mid T \in \mathcal{T}_{S_{t-1}^{\setminus E_{t-1}}} \text{ and } f(x_t) = y\}.$$

  It will be apparent from the construction that $\mathcal{T}_{S_{t-1}^{\setminus E_{t-1}}} = \mathcal{T}_0 \cup \mathcal{T}_1$ is non-empty for every t.

  – If only one of $\mathcal{T}_0, \mathcal{T}_1$ is non-empty, the adversary selects y_t such that $\mathcal{T}_{y_t}$ is non-empty.

  – Otherwise, it selects $y_t = y$ such that the probability that $\mathcal{A}$ outputs $\hat{y}_t = y$ given S_{t-1} and x_t is smallest.

  The adversary further selects some ψ_t such that $(f, \psi_t) \in \mathcal{T}_{y_t}$ for some f.
- After observing $(\hat{y}_t, \hat{x}_t)$, E_t is set as follows: If $|E_{t-1}| < k$ and $\hat{x}_t \in \mathcal{X}_i$ and $|R(S_t, \hat{x}_t, 0)| = d$ then $E_t := E_{t-1} \cup \{s\}$, where $s < t$ is the round such that $\hat{x}_t = x_s$ (note that since $i = L \geq 2$, $\hat{x}_t \in \mathcal{X}_i$ cannot have come from $H_\bullet$). Otherwise, $E_t := E_{t-1}$.

Note that since E_t never decreases, and y_t, ψ_t in round t are selected to be consistent with some teacher in $\mathcal{T}_{S_{t-1}^{\setminus E_{t-1}}}$, it follows that $\mathcal{T}_{S_t^{\setminus E_t}}$ is also non-empty. In addition, since $|E_t|$ never exceeds k and $\mathcal{T}_{S_L^{\setminus E_L}}$ is non-empty, it follows that only k-consistent sequences are generated using this adversary.

To prove a lower bound on the number of mistakes that $\mathcal{A}$ makes when interacting with V, we first prove that $\mathcal{A}$ can only have repeated a non-exception explanation for a true label of 0 for d or more times after it has made at least $(k+1)d$ mistakes.

Lemma 24 *Fix k, let $L \geq 2$ and $t \in [L]$. Let $S = S_t$ be a sequence generated by an interaction between a learner $\mathcal{A}$ and the adversary V over t rounds, with history $H_\bullet$ and exception rounds $E = E_t$. If there is an element $x_s \in S_{\mathcal{X}}^{\setminus E}$ such that $|R(S, x_s, 0)| \geq d$, then the number of prediction mistakes in S is at least $(k+1)d$.*

Proof Assume that the condition holds. Note that $x_s \in \mathcal{X}_i$. Since $x_s \in S_{\mathcal{X}}^{\setminus E}$, it follows that $s \notin E$. Let r be the first round in which $|R(S_r, x_s, 0)| = d$. Then in this round, s was not added to E_r. By the definition of the adversary, this can only happen if $|E_{r-1}| \geq k$. Thus, $|E| \geq k$. The number of mistakes in S is at least $|\cup_{s' \in E} R(S, x_{s'}, 0)|$. Moreover, $R(S, x, 0) \cap R(S, x', 0) = \emptyset$ for $x \neq x'$. Lastly, $s' \in E$ implies that $|R(S, x_{s'}, 0)| \geq d$. Therefore, the number of mistakes in S is at least $\sum_{s' \in E} |R(S, x_{s'}, 0)| + |R(S, x_s, 0)| \geq kd + d = (k+1)d$. ■

Next, we prove that as long as no non-exception explanation has been used by $\mathcal{A}$ at least d times for a true label of 0, the adversary can set the next label arbitrarily.

Lemma 25 *Fix k, let $L \geq 2$ and $t \in [L]$. Let $S = S_t$ be a sequence generated by an interaction between a learner $\mathcal{A}$ and the adversary V over t rounds, with history $H_\bullet$ and exception rounds $E = E_t$. If for all elements $x_s \in S_{\mathcal{X}}^{\setminus E}$ we have $|R(S^{\setminus E}, x_s, 0)| < d$, then for each $y \in \{0,1\}$, there is a teacher $T_y = (f_y, \psi_y) \in \mathcal{T}_{S^{\setminus E}}$ such that $f_y(x_{t+1}) = y$.*

Proof Assume that the conditions of the lemma hold. The goal is to show that for any $y^* \in \{0,1\}$ there exists a teacher $T^* = (f^*, \psi^*) \in \mathcal{T}_{S^{\setminus E}}$ such that $f^*(x_{t+1}) = y^*$. As argued above, $\mathcal{T}_{S^{\setminus E}}$ is non-empty. In addition, by the definition of the adversary, $S_{\mathcal{X}}$ includes only examples from $\mathcal{X}_i$. Also, $\mathcal{F}_i$ includes all labelings of $\mathcal{X}_i$. Thus, we can find an f^* that is consistent with any labeling of $x_1, \ldots, x_{t+1}$. Set $f^* \in \mathcal{F}_i$ such that for all $r \in [t] \setminus E$, $f^*(x_r) = y_r$, and for $r \in E$, $f^*(x_r) = 0$. In addition, $f^*(x_{t+1}) = y^*$.

We have left to show that there exists some $\psi^* \in \Psi_{f^*}$ such that for all $r \in [t] \setminus E$ and $\phi_r \neq \perp$, we have $\psi^*(x_r, \hat{x}_r) = \phi_r$. From the definition of Ψ_{f^*}, it suffices to find a function $\bar{a} : \mathcal{X}_i \times \{0,1\} \to \mathbb{F}_{p_i}^d$ such that $\psi^* = \psi_{\bar{a}}^{f^*}$ satisfies the requirement, while also $\forall \hat{x} \in \mathcal{X}_i, f^*(\hat{x}) = 1 \Rightarrow \bar{a}(\hat{x}, 0)(0) = b_i(f^*)$.

Consider a round r with $\phi_r \neq \perp$. If $\hat{x}_r$ appears in $H_\bullet$, then $\hat{x}_r \notin \mathcal{X}_i$, therefore $\phi_r = \mathbb{I}[x_r]$ regardless of the choice of teacher. Now, consider a round $r \in [t] \setminus E$ in which $\hat{x}_r \in \mathcal{X}_i$, and observe that the value of $\bar{a}$ affects $\psi_{\bar{a}}^{f^*}(x_r, \hat{x}_r)$ only via the value of $\bar{a}(\hat{x}_r, f^*(x_r)) = \bar{a}(\hat{x}_r, y_r)$. Therefore, we can set the value of $\bar{a}(\hat{x}, y)$ for each $(\hat{x}, y) \in \mathcal{X}_i \times \{0,1\}$ separately, by making sure $\psi_{\bar{a}}^{f^*}(x_r, \hat{x}_r) = \phi_r$ for rounds $r \in R(S^{\setminus E}, \hat{x}, y)$. Note that all $\hat{x}$ for which $R(S^{\setminus E}, \hat{x}, y)$ is non-empty for some $y \in \{0,1\}$ are equal to x_s for some $s \in [t]$.

To set $\bar{a}(x_s, y)$ for non empty $R(S^{\setminus E}, x_s, y)$, we distinguish between two cases:

- $s \in [t] \setminus E$ and $y = 0$. In this case, since $s \notin E$, by our assumption, $|R(S^{\setminus E}, x_s, 0)| < d$. Thus, by Lagrange's interpolation theorem, there exist coefficients $\bar{c} \in \mathbb{F}_{p_i}^d$ such that $\bar{c}(0) = P'[\bar{c}](0) = b_i(f^*)$, and for all rounds $r \in R(S^{\setminus E}, x_s, 0)$, $P[\bar{c}](x_r) = x'_r$. Set $\bar{a}(x_s, 0) := \bar{c}$. Then for all such rounds r,

$$\psi_{\bar{a}}^{f^*}(x_r, \hat{x}_r) = \psi_{\bar{a}}^{f^*}(x_r, x_s) = \mathbb{I}[x_r, P[\bar{a}(x_s, f^*(x_r))](x_r)\,] = \mathbb{I}[x_r, P[\bar{a}(x_s, 0)](x_r)\,] = \phi_r.$$

- Otherwise ($s \in E$ or $y \neq 0$). From the definition of the adversary V, there exists some teacher $T' = (f', \psi_{\bar{a}'}^{f'}) \in \mathcal{T}$ such that for all rounds $r \in [t] \setminus E$, if $\phi_r \neq \perp$ then $\phi_r = \psi_{\bar{a}'}^{f'}(x_r, \hat{x}_r) = \mathbb{I}[x_r, x'_r]$ for some $x'_r \in [2^{i+1} + 1, 2^{i+1} + p_i]$ and $f'(x_r) = y_r$. Set $\bar{a}(x_s, y) := \bar{a}'(x_s, y)$. Then

$$\begin{aligned}\psi_{\bar{a}}^{f^*}(x_r, x_s) &= \mathbb{I}[x_r, P[\bar{a}(x_s, f^*(x_r))](x_r)\,] = \mathbb{I}[x_r, P[\bar{a}(x_s, y)](x_r)\,] \\ &= \mathbb{I}[x_r, P[\bar{a}'(x_s, y)](x_r)\,] = \mathbb{I}[x_r, P[\bar{a}'(x_s, f'(x_r))](x_r)\,] = \psi_{\bar{a}'}^{f'}(x_r, x_s) = \phi_r.\end{aligned}$$

Lastly, for any $\hat{x}$ such that $R(S^{\setminus E}, \hat{x}, y)$ is empty for both $y \in \{0,1\}$, set $\bar{a}(\hat{x}, 0)(0) := b_i(f^*)$ and set all other values arbitrarily. The resulting $\psi^* = \psi_{\bar{a}}^{f^*}$ outputs ϕ_r for $(x_r, \hat{x}_r)$ for every $r \in [t] \setminus E$.

To verify that $\forall \hat{x} \in \mathcal{X}_i, f^*(\hat{x}) = 1 \Rightarrow \bar{a}(\hat{x}, 0)(0) = b_i(f^*)$, first observe that if $\hat{x} = x_s$ for some $s \in E$, then by the choice of f^*, $f^*(\hat{x}) = 0$, thus the condition does not hold. In all other cases, $\bar{a}(x_s, 0)(0) = b_i(f^*)$. The condition thus holds for $\bar{a}$, hence $\psi_{\bar{a}}^{f^*} \in \Psi_f$ and the teacher $T = (f^*, \psi_{\bar{a}}^{f^*})\mathcal{T}_{S^{\setminus E}}$ is consistent with all non-exception rounds $[t] \setminus E$. ■

We can now prove the main lower bound.

Proof [of Lemma 23] By Lemma 24 and Lemma 25 in every round t until which the learner has made $(k+1)d-1$ or fewer mistakes, for each $y \in \{0,1\}$, there is a teacher $T_y = (f_y, \psi_y) \in \mathcal{T}_{S_t^{\setminus E_t}}$ such that $f_y(x_{t+1}) = y$. In such a round $t+1$, by the definition of the adversary V, $\mathcal{A}$ makes a mistake whenever it selects the more likely label according to its own distribution in that round. The probability that this happens in any specific round is at least $1/2$. Let $n := (k+1)d$. If the number of such rounds is at least n then the number of mistakes is also at least n. The probability that there are fewer than n such rounds is at most $P_{n,L} := \sum_{j=0}^{n-1} \binom{L}{j} \left(\frac{1}{2}\right)^L$. Therefore,

$$M^L(\mathcal{A}, V, H) \geq n \cdot (1 - P_{n,L}).$$

We upper bound $P_{n,L}$ using the standard binomial upper bound (see, e.g., Shalev-Shwartz and Ben-David, 2014, Lemma A.5)

$$P_{n,L} = 2^{-L} \sum_{j=0}^{n-1} \binom{L}{j} \leq 2^{-L} \left(\frac{eL}{n-1}\right)^{(n-1)}.$$

For $L \geq 4(k+1)d > 4(n-1)$ we can further simplify

$$\begin{aligned} P_{n,L} &\leq 2^{-4(n-1)} 2^{-(L-4(n-1))} e^{n-1} \left(\frac{L}{4(n-1)}\right)^{n-1} 4^{n-1} \\ &= (4e/2^4)^{n-1} \cdot 2^{-(L-4(n-1))} \left(\frac{L}{4(n-1)}\right)^{n-1} \\ &\leq \left(\frac{3}{4}\right)^{n-1} \cdot (L^{n-1} 2^{-L}) \cdot (4(n-1))^{n-1} 2^{-4(n-1)}. \end{aligned}$$

Since $g(x) := \frac{x^a}{2^x}$ is a monotonously decreasing function in $x \geq 1$ for every $a \geq 0$ and $L \geq 4(n-1)$, we have $L^{n-1}2^{-L} \leq (4(n-1))^{n-1}2^{-4(n-1)}$. Therefore, $P_{n,L} \leq (3/4)^{n-1}$. It follows that

$$M^L(\mathcal{A}, V, H) \geq n - n(3/4)^{n-1}.$$

Since $x(3/4)^{x-1} < 2$ for every $x \geq 1$, we have $n(3/4)^{n-1} < 2$. Therefore,

$$M^L(\mathcal{A}, V, H) \geq n - 1 = (k+1)d - 1.$$

This completes the proof. ■