AISTATS 2020 Supplementary Material

Robust Learning from Discriminative Feature Feedback

Sanjoy Dasgupta and Sivan Sabato

A Deferred Proofs

Proof of Theorem 1. For (a), we first observe that we may assume without loss of generality that the components in $\mathcal{G}$ are pairwise disjoint: iteratively, for any two components G_0, G_1 that are not pairwise disjoint, replace them with G_0', G_1' such that, for $i \in \{0,1\}$,

$$G_i' := (G_i \setminus G_{1-i}) \cup \{x \in G_0 \cap G_1 \mid G(x) = G_i\}.$$

The result is a representation with the same number of components as $\mathcal{G}$ that are pairwise disjoint, and all the responses of the teacher in the interaction protocol remain the same.

Let c^* be a concept that agrees with $\bar{c}$ on all but the k exceptions, such that $|M(c^*, \mathcal{G})| = 0$. We prove the upper bound by induction on k. Suppose that for some value of k, for any concept c' such that $|M(c', \mathcal{G})| = k$, there is a representation $\mathcal{G}'$ of size $m' \leq m + dk$ that satisfies $|M(c', \mathcal{G}')| = 0$. This trivially holds for $k = 0$.

Now, consider a concept $\bar{c}$ such that $|M(\bar{c}, \mathcal{G})| = k + 1$. Let c' be a concept which agrees with c^* on all but k elements, and agrees with $\bar{c}$ on all but one element. Let $\mathcal{G}' = \{G_1', \ldots, G_{m'}'\}$ be the representation assumed by the induction hypothesis for c', and let x be the single element such that $\bar{c}(x) \neq c'(x)$. We construct a representation $\bar{\mathcal{G}}$ for $\bar{c}$.

Under the disjointness assumption, there is a single component which includes x. Suppose it is G_1'. For each $j \in [d]$, define the components $\bar{G}(j)$ as follows. Define $P_j^x := \{z \in \mathcal{X} \mid \phi_j(z) \neq \phi_j(x)\}$. Let $\bar{G}(j) := G_1' \cap P_j^x$. Define an additional singleton component $\bar{G}_x = \{x\}$. Note that $\{\bar{G}(j))\}_{j \in [d]} \cup \{\bar{G}_x\}$ exactly covers G_1'. Define

$$\bar{\mathcal{G}} := \{\bar{G}(j)\}_{j \in [d]} \cup \{G_2', \ldots, G_{m'}'\} \cup \{\bar{G}_x\}.$$

For any $\bar{G} \subseteq G_1'$ such that $\bar{G} \neq \bar{G}_x$, set $\ell(\bar{G}) := \ell(G_1')$. In addition, set $\ell(\bar{G}_x) := \bar{c}(x)$. $\bar{\mathcal{G}}$ is a legal representation, with $|M(\bar{\mathcal{G}}, \bar{c})| = 0$. The legality of $\bar{\mathcal{G}}$ can be observed by noting that the union of $\bar{\mathcal{G}}$ is $\mathcal{X}$, that the labels of all components agree with $\bar{c}$, and that any two components in $\bar{\mathcal{G}}$ with a different label can be separated by a single feature: If $\bar{G}_1 \subseteq G_i'$ and $\bar{G}_2 \subseteq G_j'$ for $i \neq j$ and their labels disagree, then the same feature that separates G_i' and G_j' separates $\bar{G}_1$ and $\bar{G}_2$. If $\bar{G}_1, \bar{G}_2 \subseteq G_1'$ and $\ell(\bar{G}_1) \neq \ell(\bar{G}_2)$, then necessarily one of the components is $\bar{G}_x$ and the other is $\bar{G}(j)$ for some j. In this case, the feature j separates the two components. The size of $\bar{\mathcal{G}}$ is $m' + d \leq m + d(k+1)$, as required by the upper bound. Note that while $\bar{\mathcal{G}}$ is not pairwise disjoint, it can be converted to a pairwise-disjoint representation by the process described above. This completes the proof of the upper bound.

To prove the lower bound (b), it suffices to consider the following example, defined over $\mathcal{X} = \{0,1\}^d$, where $\phi_j(x)$ is the value of coordinate j in x. Let $\mathcal{G} = \{\mathcal{X}\}$, $\ell(\mathcal{X}) = 0$. Let $\bar{c}$ be a concept that agrees with $c^* \equiv 0$, except on $z_0 = (0, \ldots, 0)$. Let $\mathcal{G}'$ be a representation that has $|M(\bar{c}, \mathcal{G}')| = 0$. We claim that $|\mathcal{G}'| \geq d + 1$. Consider the vectors $e_1, \ldots, e_d$. Suppose that some $G \in \mathcal{G}'$ has $e_i, e_j \in G$ for $i \neq j$. Then no single feature can separate G from the component that includes z_0. Therefore, there are at least d components for each of e_i, and a separate one for z_0. This gives a lower bound of $d + 1$. $\square$

Proof of Theorem 2. Let P_m be the set of pairs (i,j) such that $i, j \in [m]$ and $i < j$. Define a set of features $\Phi := \{\phi_{i,j}^p \mid i, j \in [m], i \neq j, p \in \{0,1\}\}$. Define a family of $2^{|P_m|}$ possible representations $\{\mathcal{G}_S\}_{S \subseteq P_m}$. The representation $\mathcal{G}_S$ includes m components $G_1, \ldots, G_m$, such that for $i < j$, component G_i is separated from component G_j using the feature $\phi_{i,j}^{S_{i,j}}$, where $S_{i,j} := \mathbb{I}[(i,j) \in S]$. In other words, for each pair of components, one of two possible features $\phi_{i,j}^0, \phi_{i,j}^1$ separates them. We further define that in G_i the separating feature is positive, while it is negative in G_j. For simplicity, we denote $\phi_{j,i} := \neg\phi_{i,j}$. Formally, G_i in representation $\mathcal{G}_S$ is the set of examples which satisfy $\left(\bigwedge_{j:i<j} \phi_{i,j}^{S_{i,j}}\right) \bigwedge \left(\bigwedge_{j:i>j} \neg\phi_{i,j}^{S_{i,j}}\right)$. In all the representations, the label of the examples in G_i is set to i.[1]

Define an example $x_{i,j}$ for $(i,j) \in P_m$ as follows: For all $l \neq i,j$ and $z \in \{0,1\}$, all the features $\phi_{i,l}^z$ and $\phi_{j,l}^z$ get the value that excludes them from G_l. The feature $\phi_{i,j}^0$ is set to positive, and $\phi_{i,j}^1$ is set to negative. Thus, in all representations S, $x_{i,j} \in G_i \cup G_j$, and $x_{i,j} \in G_i$ if and only if $(i,j) \in S$. Now, consider a stream of examples that presents $x_{i,j}$ for $(i,j) \in P_m$ in a uniformly random order and labels them using a representation $\mathcal{G}_S$ selected uniformly at random over $S \subseteq P_m$, so that the label of $x_{i,j}$ is i if $(i,j) \in S$ and j otherwise.

The stream of examples is the same for all representa-

[1] A similar example with only two labels can be shown, at the cost of a smaller multiplicative factor in the mistake bound.

tions. Thus, the only information on S can be obtained from the discriminative features. There are $\binom{m}{2}$ possible elements in S, and each discriminative feature feedback in this problem reveals whether $(i,j) \in S$ for a single pair (i,j). Moreover, if this is unknown for some pair (i,j) when $x_{i,j}$ is revealed, then both values of $S_{i,j}$ are equally likely conditioned on the run so far. In this case, any algorithm will provide the wrong label with a probability at least a half. Now, after less than $|P_m|/2$ mistakes, there is a probability of at least a half to observe such an example in the next iteration. Therefore, in the first $|P_m|/2$ examples of the stream, there is a probability of at least $1/4$ that the algorithm makes a mistake on the next example. Thus, the expected number of mistakes is at least $|P_m|/8 = \Omega(m^2)$. $\square$

To prove Lemma 13, we use the following concentration inequality.

Lemma 15. *Let $\delta \in (0, 1/e^2)$, let k be an integer and let $p \in [\frac{1}{2}, 1)$. The probability that a sum of k independent geometric random variables with probability of success p is larger than $\frac{1}{p}\min(2k\log(1/\delta), (k + 4\sqrt{k}\log^{3/2}(1/\delta)))$ is at most δ.*

Proof. This lemma follows from Hoeffding's inequality, by noting that the number of successes in N experiments with success probability p is distributed as $\mathrm{Binom}(N,p)$, and having

$$\mathbb{P}[\mathrm{Binom}(N,p) < k] \leq \exp(-2N(p - k/N)^2).$$

First, defining $N_1 := 2k\log(1/\delta)/p$, we have

$$k/N_1 = p/(2\log(1/\delta)) \leq p(1 - 1/\sqrt{2}).$$

Hence, $p - k/N_1 \geq p/\sqrt{2}$. It follows that

$$\begin{aligned}\exp(-2N_1(p - k/N_1)^2) &\leq \exp(-N_1 p^2) \\ &\leq \exp(-N_1 p/2) = \exp(-k\log(1/\delta)) \leq \delta.\end{aligned}$$

Second, suppose that $k \geq 4\log(1/\delta)$, and let $\alpha := \sqrt{\log(1/\delta)/4k} \leq \frac{1}{4}$. Defining

$$N_2 := 2(1 + 4\alpha)k/p = \frac{1}{p}(2k + 4\sqrt{k\log(1/\delta)}),$$

we have that

$$1/(p - \alpha) = 1/p + \alpha/(p(p - \alpha)) \leq (1 + 4\alpha)/p,$$

where the last inequality follows since $p \geq \frac{1}{2}$ and $\alpha \leq \frac{1}{4}$. Therefore, $N_2 \geq k/(p - \alpha)$, hence $k/N_2 \leq p - \alpha$, hence

$$\begin{aligned}\exp(-2N_2(p - k/N_2)^2) &\leq \exp(-4(k/p)\alpha^2) \\ &= \exp(-\log(1/\delta)/p) \leq \delta.\end{aligned}$$

The proof is completed by observing that the first bound in the statement of the lemma is N_1, and the second bound is always larger than N_2, and for $k \leq 4\log(1/\delta)$, it is larger than N_1. $\square$

We now prove Lemma 13.

Proof of Lemma 13. Denote by L_t the set of rules L at the end of round t of the run of `StRoDFF`. Let

$$\mathcal{L}_t = \{x \in \mathcal{X} \mid \exists C \in L_t \text{ such that } x \text{ satisfies } C\},$$

and denote $p_t := \mathbb{P}[X \in \mathcal{L}_t]$, where X is a random example drawn according to the distribution creating the input stream. We now prove the main claim: that with a high probability, a rule is not created by `StRoDFF` at round t unless $p_{t-1} \leq 1 - 2\epsilon$. The claim is proved by induction on the sequence of rules created by `StRoDFF`. For the basis of the induction, observe that $p_0 = 0$, since L_0 is empty. Therefore, the first rule created by `StRoDFF` certainly satisfies the claim for any $\epsilon < \frac{1}{2}$. For the induction step, suppose that the claim holds for the first l rules created by `StRoDFF`. Let t_0 be the round in which the l'th rule was created, and condition on the stream prefix ending in t_0. We show that the next rule also satisfies the claim.

First, for any round $t \geq t_0$ until a new rule is created, p_t is monotonic non-increasing. This is because the possible transformations, other than creating a new rule, are to restrict a rule or to delete a rule, both of which can never increase the set of examples covered by L. Therefore, if $p_{t_0} \leq 1 - 2\epsilon$, then regardless of the round t in which the next rule is created, it satisfies $p_{t-1} \leq 1 - 2\epsilon$. Thus, assume below that $p_{t_0} > 1 - 2\epsilon \geq \frac{1}{2}$. p_{t_0} is the probability that a random example observed immediately after round t_0 is satisfied by some rule in $\mathcal{L}_{t_0}$. Now, consider the first round after t_0 that an example in $\mathcal{L}_{t_0}$ arrives. Denote this round t_1. The value $T_1 := t_1 - t_0$ is a geometric random variable with a success probability p_{t_0}. By Lemma 15 with $k := 1$, $p := p_{t_0}$, with a probability at least $1 - \delta/(8t_0^2)$,

$$T_1 \leq \frac{1}{p_{t_0}}(1 + 4\log^{3/2}(8t_0^2/\delta))) < \gamma(\epsilon, 1, t_0).$$

In the last inequality we used $p_0 > 1 - 2\epsilon$ and the definition of γ. Assume below that this event holds.

Now, consider N_{lr}, which counts in `StRoDFF` the number of examples since the creation of the last rule, for which the default prediction (x_0, y_0) was provided. These are the examples that were not satisfied by any rule in L when they appeared. We prove by induction on the rounds that a new rule is not created at least until round t_1. If a new rule was not created until round $t \in \{t_0 + 1, \ldots, t_1 - 1\}$, then $L_t = L_{t_0}$ (since the set of rules does not change until t_1 when an example falls in

$\mathcal{L}_{t_0}$). In addition, $N_{lr} = t - t_0$, since the examples until round t_1 are not in $\mathcal{L}_t = \mathcal{L}_{t_0}$, thus they get the default prediction. Therefore, $t - t_0 - N_{lr} = 0$. It follows that in round t,

$$N_{lr} \le T_1 < \gamma(\epsilon, 1, t_0) \le \gamma(\epsilon, t - t_{\mathrm{lr}} - N_{lr} + 1, t).$$

This means that the condition in line 21 does not hold. Thus, under the event above, a new rule will not be created at round t. Since this holds by induction for all $t \in \{t_0 + 1, \ldots, t_1 - 1\}$, it follows that if $p_0 > 1 - 2\epsilon$ then a new rule is not created at least until the first example in $\mathcal{L}_{t_0}$ arrives.

Now, $\mathcal{L}_{t_1}$ is the set of rules after this example arrives, and the probability mass of examples in $\mathcal{L}_{t_1}$ is p_{t_1}. More generally, let t_i be the first round after t_{i-1} in which an example in $\mathcal{L}_{t_{i-1}}$ appears. If no new rule is created between t_0 and t_i, then in round t_i, the set of rules changes from $L_{t_{i-1}}$ to L_{t_i}. The number of rounds $T_i := t_i - t_{i-1}$ between each two such examples is a geometric random variable with success probability $p_{t_{i-1}}$. Let r be the number of examples satisfied by L which appear in the stream until the next rule after t_0 is created, and suppose for contradiction that $p_{t_r} > 1 - 2\epsilon$. For $q \le r$, define the random variable $S_q := \sum_{i=1}^{q} T_i$. This is a sum of q independent geometric random variables, each with a probability of success larger than $1 - 2\epsilon$ (since $p_{t_q} \ge p_{t_r}$ for all $q \le r$). Thus, S_q is dominated by a sum of independent geometric random variables with a success probability of $1 - 2\epsilon$. Therefore, by Lemma 15, with a probability at least $\delta/(8(t_0 + q - 1)^2))$,

$$\begin{aligned} S_r &\le \frac{1}{1-2\epsilon}(q + 4\sqrt{q}\log^{3/2}(8(t_0 + q - 1)^2/\delta)) \\ &< \gamma(\epsilon, q, t_0 + q - 1) + q - 1. \end{aligned}$$

Assume below that this event holds for all $q \le r$. We now prove that under the assumption on p_{t_r}, a new rule is not created until t_r, which is a contradiction. Suppose for induction that since round t_0 until round $t \le t_r - 1$, a new rule was not created. Let $q \le r$ such that $t \in \{t_{q-1} + 1, \ldots, t_q - 1\}$. We have $t_q = t_0 + S_q$. Therefore, at round t, $N_{lr} = t - t_0 - (q-1) < S_q - (q-1)$. It follows that under the assumed event, in round t

$$N_{lr} < \gamma(\epsilon, q, t_0 + q - 1) \le \gamma(\epsilon, t - t_{\mathrm{lr}} - N_{lr} + 1, t).$$

Here, we used the fact that $t_0 + q - 1 \le t$. It follows that the condition in line 21 does not hold in round t, thus a new rule is not created in this round. By induction, this holds for all $t \le t_r - 1$, which contradicts the assumption that a rule was created until round t_r. Thus, if $p_{t_r} > 1 - 2\epsilon$ then a new rule is not created at least until round t_r. Since this analysis holds for any value of r, we conclude that if all the events above hold simultaneously, then a new rule is never created in round t unless $p_{t-1} \le 1 - 2\epsilon$. By a union bound on the created rules and the sequence of examples between rule-creations, this is true with a probability at least $1 - \delta/4$. □

Proof of Theorem 9. First, we upper bound the number of mistakes on examples that are not satisfied by any rule when they are observed. Let $t_1, t_2, \ldots, t_R$, which sum to n, be the lengths of times between creations of new rules (where t_1 is time of the first rule and t_R is the time between the last rule and the end of the stream). We have by Lemma 14 that $R \le R(m, \delta) + 1$. We have $1/(1 - 2\epsilon) = 1 + 2\epsilon/(1 - 2\epsilon) \le 1 + 4\epsilon$, where the last inequality follows since $\epsilon \le \frac{1}{4}$. Hence,

$$\begin{aligned} \gamma(\epsilon, r, t) &\equiv \frac{1}{1-2\epsilon}(r + 4\sqrt{r}\log^{3/2}(8t^2/\delta)) - r + 1 \\ &\le 8\epsilon r + 8\sqrt{r}\log^{3/2}(8t^2/\delta). \end{aligned}$$

The number of mistakes resulting from examples not satisfied by any rule is upper-bounded by

$$\begin{aligned} \sum_{i=1}^{R} \gamma(\epsilon, t_i, n) &\le 8\epsilon n + 8\sum_{i=1}^{R}\sqrt{t_i}\log^{3/2}(8n^2/\delta) \\ &\le 8\epsilon n + 8\sqrt{Rn}\log^{3/2}(8n^2/\delta). \end{aligned}$$

In addition, any existing rule may generate at most $(m-1)(q(\sigma, n) + 2) + q(\epsilon, n) + 1$ mistakes (since it would be deleted after that). Note that $R = O(m\log(1/\delta))$, and $q(\epsilon, n) = O(\epsilon n + \log(n/\delta) + \sqrt{n\log(n/\delta)})$. The total upper bound is thus $O\Big(\epsilon n + \sqrt{mn}\log^2(n/\delta) + m\log(1/\delta)(\epsilon n + m(\sigma n + \log(n/\delta) + \sqrt{n\log(n/\delta)}\Big)$. Dividing by n and reorganizing, we get the error rate in the statement of the lemma. □