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 the labels of all components agree with $\bar{\mathcal{G}}$ with any two components in $\bar{\mathcal{G}}$, and their labels disagree, then the same

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 $\phi_{i,j}^0$, 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 $\phi_{i,j} := -\phi_{i,j}^1$. Formally, $G_i$ in representation $\mathcal{G}_S$ is the set of examples which satisfy $\left(\bigwedge_{1 \leq i < j \leq m} \phi_{i,j}^{S_{i,j}}\right)$ and $\left(\bigwedge_{1 \leq i \leq j \leq m} -\phi_{i,j}^{S_{i,j}}\right)$. In all the representations, the label of the examples in $G_i$ is set to $i$.

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 $\phi_{i,j}^0, \phi_{i,j}^1$ 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 representatives.
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 $|P_m|/8 = \Omega(m^2)$. \hfill \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 $p \in \left[\frac{1}{2}, 1\right]$. 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 $\delta$.

**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 Binom$(N,p)$, and having

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

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

$$\exp(-2N_1(p-k/N_1)^2) \leq \exp(-N_1p^2) \leq \exp(-N_1p/2) = \exp(-k \log(1/\delta)) \leq \delta.$$

Second, suppose that $k \geq 4\log(1/\delta)$, and let $\alpha := \sqrt{\log(1/\delta)/4k} \leq 1/4$. 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 $p \geq 1/2$ and $\alpha \leq 1/4$. Therefore, $N_2 \geq k/(p-\alpha)$, hence $k/N_2 \leq p - \alpha$, hence

$$\exp(-2N_2(p-k/N_2)^2) \leq \exp(-4(k/p)^2(\alpha^2)) \leq \exp(-\log(1/\delta)/p) \leq \delta.$$
\( \mathcal{L}_{t_0} \). In addition, \( N_{t_F} = t - t_0 \), since the examples until round \( t_1 \) are not in \( \mathcal{L}_{t} = \mathcal{L}_{t_0} \), thus they get the default prediction. Therefore, \( t - t_0 - N_{t_F} = 0 \). It follows that in round \( t \),

\[
N_{t_F} \leq T_1 < \gamma(\epsilon, 1, t_0) \leq \gamma(\epsilon, t - t_{1_F} - N_{t_F} + 1, t).
\]

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 \( \mathcal{L}_{t_{i-1}} \) to \( \mathcal{L}_{t_i} \). 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 \( q \leq r \), 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 \( p_{t_i} \geq p_{t_r} \) for all \( q \leq r \)). 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 \( \delta/(8(t_0 + q - 1)^2) \),

\[
S_r \leq \frac{1}{1 - 2\epsilon} (q + 4\sqrt{q} \log^{3/2}(8(t_0 + q - 1)^2)) \leq \gamma(\epsilon, q, t_0 + q - 1) + q - 1.
\]

Assume below that this event holds for all \( q \leq r \). 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 \( t \leq t_r - 1 \), a new rule was not created. Let \( q \leq r \) such that \( t \in \{ t_{q-1} + 1, \ldots, t_q - 1 \} \). We have \( t_q = t_0 + S_q \). Therefore, at round \( t \), \( N_{t_r} = t - t_0 - (q - 1) < S_q - (q - 1) \). It follows that under the assumed event, in round \( t \)

\[
N_{t_r} < \gamma(\epsilon, q, t_0 + q - 1) \leq \gamma(\epsilon, t - t_{1_F} - N_{t_F} + 1, t).
\]

Here, we used the fact that \( t_0 + q - 1 \leq t \). 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 \( t \leq t_r - 1 \), 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 \( p_{t-1} \leq 1 - 2\epsilon \). By a union bound on the created rules and the sequence of examples between rule-creations, this is true with a probability at least \( 1 - \delta/4 \).

\[ \square \]

**Proof of Theorem 12** 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 \( R \leq R(m, \delta) + 1 \). We have \( 1/(1 - 2\epsilon) = 1 + 2\epsilon/(1 - 2\epsilon) \leq 1 + 4\epsilon \), where the last inequality follows since \( \epsilon \leq \frac{1}{4} \). Hence,

\[
\gamma(\epsilon, r, t) \equiv \frac{1}{1 - 2\epsilon} (r + 4\sqrt{r} \log^{3/2}(8t^2/\delta)) - r + 1 \leq 8\epsilon r + 8\sqrt{r} \log^{3/2}(8t^2/\delta).
\]

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

\[
\sum_{i=1}^{R} \gamma(\epsilon, t_i, n) \leq 8\epsilon n + 8\sum_{i=1}^{R} \sqrt{t_i} \log^{3/2}(8n^2/\delta)
\]

\[
\leq 8\epsilon n + 8\sqrt{n} \log^{3/2}(8n^2/\delta).
\]

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 \( q(\epsilon, n) = O(n \log(n/\delta) + \sqrt{n} \log(n/\delta)) \). The total upper bound is thus \( O(\epsilon n + \sqrt{mn} \log^2(n/\delta) + m \log(1/\delta)(en + m(\sigma n + \log(n/\delta) + \sqrt{n} \log(n/\delta))) \). Dividing by \( n \) and reorganizing, we get the error rate in the statement of the lemma.

\[ \square \]