9 Appendix

9.1 Guarantees for General Agnostic Algorithm

In this section, we give proofs for the guarantees of Algorithm 1. We begin with some definitions, starting with how empirical loss estimates are made.

Definition 1. *Given a hypothesis $h \in \mathcal{H}$, and a set of pairs $\mathcal{S} = \{(x_i, y_i) : x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}_{i=1}^N$, let*

$$L_{\mathcal{S}}(h) := \frac{1}{N}\left(\sum_{i=1}^{N} \mathbb{1}[h(x_i) \neq y_i]\right)$$

the standard empirical loss of h on $\mathcal{S}$. Let $L_\emptyset(h) := 1$.

The convention to let $L_\emptyset(h) = 1$ allows us to "collapse" the two-part loss estimates in the case the probability of drawing an unlabeled sample in a specific region is 0; under the specification of the algorithm, $\mathcal{S} = \emptyset$ if and only if the probability of a sample falling in the disagreement region or its complement is 0 under D_g, in which case we can safely ignore estimation in one of these regions.

Definition 2. *Given a set of classifiers $\mathcal{H}' \subseteq \mathcal{H}$, we say "$\mathcal{H}'$ agrees on a subset $S \subseteq \mathcal{X}$" if for each $x \in S$ and for each pair $(h, h') \in \mathcal{H}' \times \mathcal{H}'$, it holds that $h(x) = h'(x)$.*

We now recall the two-part estimator for the loss of a hypothesis introduced above.

Definition 3. *Fix a group distribution D_g, some $\mathcal{H}' \subseteq \mathcal{H}$, a hypothesis $h \in \mathcal{H}'$, and some $R \subseteq \mathcal{X}$ which is measurable with respect to each marginal of D_g and for which $\mathcal{H}'$ agrees on R^c. Given sets of pairs $\mathcal{S}_{R,g}$ and $\mathcal{S}_{R^c,g}$, and some arbitrarily chosen classifier $h_{\mathcal{H}'} \in \mathcal{H}'$, let*

$$L_{\mathcal{S};R}(h \mid g) := \mathbb{P}_{D_g}(x \in R) \cdot L_{\mathcal{S}_{R,g}}(h) + \mathbb{P}_{D_g}(x \in R^c) \cdot L_{\mathcal{S}_{R^c,g}}(h_{\mathcal{H}'}).$$

As mentioned in the main body, $h_{\mathcal{H}'}$ must be used in the estimate of the loss under D_g in the "agreement region" for all $h \in \mathcal{H}$. The extent to which this estimator is useful can be captured by standard uniform convergence arguments. To this end, we first introduce a function that will prove to control its deviations nicely.

Definition 4. *Given a confidence parameter $\delta \in (0,1)$, a group distribution $D_g \in \mathcal{G}$, some $R \subseteq \mathcal{X}$ that is measurable with respect to each marginal $D_g \in \mathcal{G}$, and sample sizes $m, m' > 0$, define the function*

$$\Gamma_g(\delta, R, m, m') := \begin{cases} \mathbb{P}_{D_g}(x \in R)\left(\frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(2em/d)}{m}}\right) + \sqrt{\frac{\ln(4/\delta)}{2m'}} \\ \qquad \textit{if } \mathbb{P}_{D_g}(x \in R) > 0, \mathbb{P}_{D_g}(x \in R^c) > 0 \\ \frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(2em/d)}{m}} \\ \qquad \textit{if } \mathbb{P}_{D_g}(x \in R) > 0, \mathbb{P}_{D_g}(x \in R^c) = 0 \\ \sqrt{\frac{\ln(4/\delta)}{2m'}} \\ \qquad \textit{if } \mathbb{P}_{D_g}(x \in R) = 0, \mathbb{P}_{D_g}(x \in R^c) > 0. \end{cases}$$

Lemma 1. *Fix $\delta \in (0,1)$, a set of group distributions $\mathcal{G}$, and a group distribution $D_g \in \mathcal{G}$ arbitrarily. Further, fix a subset $R \subseteq \mathcal{X}$ measurable with respect to each marginal of $D_g \in \mathcal{G}$, and a set of classifiers $\mathcal{H}' \subseteq \mathcal{H}$ with the property that $\mathcal{H}'$ agree on R^c. Suppose we query $m > 0$ unlabeled samples from $U_g(R)$, and $m' > 0$ samples from $U_g(R^c)$. Suppose further that we label the output via calls to $O_g(\cdot)$, forming the labeled samples $\mathcal{S}_{R,g}$ and $\mathcal{S}_{R^c,g}$, respectively; if either $\mathbb{P}_{D_g}(x \in R) = 0$ or $\mathbb{P}_{D_g}(x \in R^c)$, then we set the corresponding sample to be $\emptyset$. Then with probability $\geq 1 - \delta$, it holds for all $h \in \mathcal{H}'$ that*

$$|L_{\mathcal{G}}(h \mid g) - L_{\mathcal{S};R}(h \mid g)| \leq \Gamma_g(\delta, R, m, m').$$

Further, for all $\gamma > 0$, if $m \geq \frac{16(\mathbb{P}_{D_g}(x \in R))^2}{\gamma^2}(2d\ln(8/\gamma) + \ln(8/\delta))$ and $m' \geq \frac{2\ln(4/\delta)}{\gamma^2}$, then $\Gamma_g(\delta, R, m, m') < \gamma$.

Proof. We begin with the case where both $\mathbb{P}_{D_g}(x \in R) \neq 0$ and $\mathbb{P}_{D_g}(x \in R^c) \neq 0$. In this case, we are able to draw unlabeled samples from both regions, and neither $\mathcal{S}_{R,g}$ nor $\mathcal{S}_{R,g}$ is $\emptyset$.

By a lemma of Vapnik [28], we have that with probability $\geq 1 - \delta/2$ over the draw of m samples from $U_g(R)$ and their labeling via $O_g(\cdot)$, that simultaneously for each $h \in \mathcal{H}'$:

$$\left| L_{\mathcal{S}_{R,g}}(h) - \mathbb{P}_{D_g}\left(h(x) \neq y \middle| x \in R\right) \right| \leq \frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(2em/d)}{m}}.$$

In R^c, all $h \in \mathcal{H}'$ agree, and so estimating the conditional loss for each $h \in \mathcal{H}'$ in this region is as statistically hard as estimating a single Bernoulli parameter, which we do by arbitrarily choosing a classifier to use for the loss estimate in this part of space. Thus, by definition of the two-part estimator and Hoeffding's inequality [29], we have with probability $\geq 1 - \delta/2$ for all $h \in \mathcal{H}'$ simultaneously

$$\left| L_{\mathcal{S}_{R^c,g}}(h_{\mathcal{H}'}) - \mathbb{P}_{D_g}\left(h(x) \neq y \middle| x \in R^c\right) \right| \leq \sqrt{\frac{\ln(4/\delta)}{2m'}}.$$

By a union bound, with probability $\geq 1 - \delta$, both of these events take place, and so for all $h \in \mathcal{H}'$ simultaneously,

$$\begin{aligned}
L_{\mathcal{G}}(h \mid g) &= \mathbb{P}_{D_g}\big(h(x) \neq y \mid x \in R\big) \cdot \mathbb{P}_{D_g}(x \in R) \\
&\qquad + \mathbb{P}_{D_g}\left(h(x) \neq y \mid x \in R^c\right) \cdot \mathbb{P}_{D_g}(x \in R^c) \\
&\leq \left(L_{\mathcal{S}_{R,g}}(h) + \sqrt{\left(\ln(8/\delta) + d\ln(2em/d)\right)/m} \right) \cdot \mathbb{P}_{D_g}(x \in R) \\
&\qquad + \left(L_{\mathcal{S}_{R^c,g}}(h_{\mathcal{H}'}) + \sqrt{\ln(4/\delta)/2m'} \right) \cdot \mathbb{P}_{D_g}(x \in R^c) \\
&\leq L_{\mathcal{S};R}(h \mid g) + \Gamma_g(\delta, R, m, m').
\end{aligned}$$

The lower bound leading to the absolute value is analogous. Vapnik [28] also tells us that for any $\gamma' > 0$, a sample of size $m \geq \frac{4}{\gamma'^2}\left(2d\ln(4/\gamma') + \ln(8/\delta)\right)$ is sufficient to yield

$$\sqrt{\left(\ln(8/\delta) + d\ln(2em/d)\right)/m} < \gamma'.$$

Let $\gamma' = \gamma/2\mathbb{P}_{D_g}(x \in R)$. Thus, substituting for γ' and bounding the probability inside the natural log above by 1,

$$m \geq \mathbb{P}_{D_g}(x \in R)^2 \frac{16}{\gamma^2}\left(2d\ln(8/\gamma) + \ln(8/\delta)\right)$$

implies that

$$\frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(2em/d)}{m}} < \frac{\gamma}{2\mathbb{P}_{D_g}(x \in R)}.$$

As a corollary to Hoeffding, if $m' \geq 2\ln(4/\delta)/\gamma^2$, then $\sqrt{\log(4/\delta)/2m'} < \gamma/2$. Thus, we may write

$$\Gamma_g(\delta, R, m, m') = \mathbb{P}_{D_g}(x \in R)\left(\frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(2em/d)}{m}}\right) + \sqrt{\frac{\ln(4/\delta)}{2m'}} < \gamma/2 + \gamma/2 = \gamma.$$

Now suppose that $\mathbb{P}_{D_g}(x \in R^c) = 0$. In this case, we have $\mathcal{S}_{R^c,g} = \emptyset$. Again, we have that with probability $\geq 1 - \delta/2$,

$$\left| L_{\mathcal{S}_{R,g}}(h) - \mathbb{P}_{D_g}\left(h(x) \neq y \middle| x \in R\right) \right| \leq \frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(2em/d)}{m}}.$$

When $\mathbb{P}_{D_g}(x \in R^c) = 0$, it holds that $\mathbb{P}_{D_g}(x \in R) = 1$, and so

$$\begin{aligned}
L_{\mathcal{G}}(h \mid g) &= \mathbb{P}_{D_g}\left(h(x) \neq y \mid x \in R\right) \cdot \mathbb{P}_{D_g}(x \in R) \\
&\qquad + \mathbb{P}_{D_g}\left(h(x) \neq y \mid x \in R^c\right) \cdot \mathbb{P}_{D_g}(x \in R^c) \\
&= \mathbb{P}_{D_g}\left(h(x) \neq y \mid x \in R\right) \\
&\leq L_{\mathcal{S}_{R,g}}(h) + \sqrt{\left(\ln(8/\delta) + d\ln(2em/d)\right)/m} \\
&= L_{\mathcal{S};R}(h \mid g) + \Gamma_g(\delta, R, m, m'),
\end{aligned}$$

where the final equality comes from fact that $\mathbb{P}_{D_g}(x \in R^c) = 0$ and $\mathbb{P}_{D_g}(x \in R) = 1$, as well as the definitions of $L_{\mathcal{S};R}(h \mid g)$ and $\Gamma_g(\delta, R, m, m')$. Similarly to the above, if we let $\gamma' = \gamma/2\mathbb{P}_{D_g}(x \in R) = \gamma/2$, then

$$m \geq \frac{16}{\gamma^2}\left(2d\ln(8/\gamma) + \ln(8/\delta)\right)$$

implies that

$$\frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(2em/d)}{m}} < \frac{\gamma}{2},$$

which by the definition of $\Gamma_g(\delta, R, m, m')$ when $\mathbb{P}_{D_g}(x \in R^c) = 0$ gives us $\Gamma_g(\delta, R, m, m') < \gamma/2 < \gamma$. The case where $\mathbb{P}_{D_g}(x \in R) = 0$ follows the previous argument for when $\mathbb{P}_{D_g}(x \in R^c) = 0$.

□

Definition 5. *Given a collection of group distributions $\mathcal{G}$, some $\mathcal{H}' \subseteq \mathcal{H}$, a hypothesis $h \in \mathcal{H}'$, some subset $R \subseteq \mathcal{X}$ measurable with respect to each marginal of $D_g \in \mathcal{G}$, and labeled samples $\mathcal{S}_{R,k}$ and $\mathcal{S}_{R^c,k}$, we define the empirical estimate of the multi-group loss of h parameterized by R via*

$$L^{\max}_{\mathcal{S};R}(h) := \max_{g \in [G]} L_{\mathcal{S};R}(h \mid g).$$

Having recalled the way in which we form empirical estimates for the group worst-case loss of a given hypothesis, we can show a simple concentration lemma for this group worst-case loss estimator using the concentration property for individual groups proved in Lemma 1,

Lemma 2. *Fix $\delta \in (0,1)$, a set of group distributions $\mathcal{G}$, a subset $R \subseteq \mathcal{X}$ measurable with respect to each marginal of $D_g \in \mathcal{G}$, and a set of classifiers $\mathcal{H}' \subseteq \mathcal{H}$ that agree on R^c. Suppose for each $g \in [G]$, we query $m_g > 0$ unlabeled samples from $U_g(R)$, and $m'_g > 0$ samples from $U_g(R^c)$. Suppose further that we label the outputs via calls to $O_g(\cdot)$, forming the labeled samples $\mathcal{S}_{R,g}$ and $\mathcal{S}_{R^c,g}$, respectively, for each $g \in [G]$; if $\mathbb{P}_{D_g}(x \in R) = 0$ or $\mathbb{P}_{D_g}(x \in R^c) = 0$, then we set the corresponding sample to be $\emptyset$. Then with probability $\geq 1 - \delta$, it holds for all $h \in \mathcal{H}'$ that*

$$\left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{S};R}(h)\right| \leq \max_{g' \in [G]} \Gamma_{g'}(\delta/G, m_{g'}, m'_{g'}).$$

Proof. By Lemma 1 and a union bound, it holds with probability $\geq 1 - \delta$ that on all D_g, for all $h \in \mathcal{H}'$ simultaneously, that

$$|L_{\mathcal{G}}(h \mid g) - L_{\mathcal{S};R}(h \mid g)| \leq \Gamma_g(\delta/G, m_g, m'_g).$$

Thus we may write

$$\begin{aligned}
\left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{S};R}(h)\right| &= \left|\max_{g' \in [G]} L_{\mathcal{G}}(h \mid g') - \max_{g' \in [G]} L_{\mathcal{S};R}(h \mid g)\right| \\
&\leq \max_{g' \in [G]} \left|L_{\mathcal{G}}(h \mid g') - L_{\mathcal{S};R}(h \mid g')\right| \\
&\leq \max_{g' \in [G]} \Gamma_{g'}(\delta/G, m_{g'}, m'_{g'}).
\end{aligned}$$

□

We now use Lemma 2 to show that Algorithm 1 is conservative enough that the optimal hypothesis h^* is never eliminated from contention throughout the run of the algorithm with high probability.

Lemma 3. *Fix $\delta \in (0,1)$, a collection of group distributions $\mathcal{G}$, and a hypothesis class $\mathcal{H}$ with $d < \infty$ arbitrarily. With probability $\geq 1 - \delta$, it holds after each iteration i of Algorithm 1 that $h^* \in \mathcal{H}_{i+1}$.*

Proof. By Lemmas 1 and 2, and a union bound over iterations, the number of samples labeled at each iteration is sufficient for us to conclude that with probability $\geq 1 - \delta$, for for every iteration i

and for each $h \in \mathcal{H}_i$, it holds that[2]

$$|L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{G}}(h)| \leq 2^{I-i}\epsilon/8.$$

We give an inductive argument conditioned on this high probability event. When $i = 1$, we have $h^* \in \mathcal{H}_1$ because $\mathcal{H}_1 = \mathcal{H}$, and $h^* \in \mathcal{H}$ by definition. If $h^* \in \mathcal{H}_i$ for $i \geq 1$, then $h^* \in \mathcal{H}_{i+1}$ if and only if

$$L^{\max}_{\mathcal{S};R_i}(h^*) \leq L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) + 2^{I-i}\epsilon/4.$$

When for each $h \in \mathcal{H}_i$, it holds that $|L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{G}}(h)| \leq 2^{I-i}\epsilon/8$, we may write

$$\begin{aligned}
L^{\max}_{\mathcal{S};R_i}(h^*) - L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) &\leq L^{\max}_{\mathcal{S};R_i}(h^*) - L^{\max}_{\mathcal{G}}(h^*) + L^{\max}_{\mathcal{G}}(\hat{h}_i) - L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) \\
&\leq \left|L^{\max}_{\mathcal{S};R_i}(h^*) - L^{\max}_{\mathcal{G}}(h^*)\right| + \left|L^{\max}_{\mathcal{G}}(\hat{h}_i) - L^{\max}_{\mathcal{S};R_i}(\hat{h}_i)\right| \\
&\leq 2^{I-i}\epsilon/8 + 2^{I-i}\epsilon/8 \\
&= 2^{I-i}\epsilon/4,
\end{aligned}$$

where the first inequality comes from the optimality of h^*. Thus, we must have $h \in \mathcal{H}_{i+1}$. □

Now, using the fact that the optimal hypothesis stays in contention throughout the run of the algorithm, we can give a guarantee on the true error of each hypothesis $h \in \mathcal{H}_{i+1}$. The idea is that using concentration and the small empirical error of each $h \in \mathcal{H}_{i+1}$, we can say that the true errors of each $h \in \mathcal{H}_{i+1}$ are similar to the true errors of the ERM hypothesis $\hat{h}_i$, and then use the true error of $\hat{h}_i$ as a reference point to which we can compare the true error of $h \in \mathcal{H}_{i+1}$ and h^*.

Lemma 4. *Fix $\delta \in (0, 1)$, a collection of group distributions $\mathcal{G}$, and a hypothesis class $\mathcal{H}$ with $d < \infty$ arbitrarily. Then with probability $\geq 1 - \delta$, after every iteration i of Algorithm 1, it holds for all $h \in \mathcal{H}_{i+1}$ that*

$$\left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{G}}(h^*)\right| \leq 2^{I-i}\epsilon.$$

Proof. If $h \in \mathcal{H}_{i+1}$, then by the specification of the algorithm, it holds that

$$L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) \leq 2^{I-i}\epsilon/4.$$

Because $\hat{h}_i$ is the ERM hypothesis at iteration i, it holds that $L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) - L^{\max}_{\mathcal{S};R_i}(h) \leq 0 < 2^{I-i}\epsilon/4$, and thus we may conclude

$$\left|L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{S};R_i}(\hat{h}_i)\right| \leq 2^{I-i}\epsilon/4.$$

By Lemma 2 and the number of samples labeled at each iteration, with probability $\geq 1 - \delta$, it holds for all iterations and for all $h \in \mathcal{H}_i$ that

$$\left|L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{G}}(h)\right| \leq 2^{I-i}\epsilon/8.$$

Conditioned on this event, if $h \in \mathcal{H}_{i+1}$, we have

$$\begin{aligned}
\left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{G}}(\hat{h}_i)\right| &= \left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{S};R_i}(h) + L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) + L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) - L^{\max}_{\mathcal{G}}(\hat{h}_i)\right| \\
&\leq \left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{S};R_i}(h)\right| + \left|L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{S};R_i}(\hat{h}_i)\right| + \left|L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) - L^{\max}_{\mathcal{G}}(\hat{h}_i)\right| \\
&\leq 2^{I-i}\epsilon/8 + 2^{I-i}\epsilon/4 + 2^{I-i}\epsilon/8 \\
&= 2^{I-i}\epsilon/2.
\end{aligned}$$

By Lemma 3, it holds that $h^* \in \mathcal{H}_{i+1}$ whenever $\left|L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{G}}(h)\right| \leq 2^{I-i}\epsilon/8$ for all $h \in \mathcal{H}_i$ at all iterations. Thus, this bound on the true error difference with the ERM $\hat{h}_i$ applies to h^*, and we may write for arbitrary $h \in \mathcal{H}_{i+1}$ that

$$\left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{G}}(h^*)\right| \leq \left|L^{\max}_{\mathcal{G}}(h) - L^{\max}_{\mathcal{G}}(\hat{h}_i)\right| + \left|L^{\max}_{\mathcal{G}}(\hat{h}_i) - L^{\max}_{\mathcal{G}}(h^*)\right| \leq 2^{I-i}\epsilon,$$

which is the desired result. □

[2] We do not directly apply Lemma 1 with $\gamma = \epsilon 2^{I-i}/8$ here. We use this quantity in the outer dependence on γ of Lemma 1, but for the natural log dependence on γ, we sub in $\epsilon/8$ to simplify the analysis. Thus we take slightly more samples than Lemma 1 directly suggests. Because we take the largest probability of the disagreement region over groups as m_i, it holds that m_g is at the smallest the sample size suggested by Lemma 1 for each g.

Definition 6. *Given a group distribution $D_g \in \mathcal{G}$, a hypothesis $h \in \mathcal{H}$, and a radius $r \geq 0$, let the "D_g - disagreement ball in $\mathcal{H}$ of radius r about h" be*

$$B_g(h, r) := \{h' \in \mathcal{H} : \rho_g(h, h') \leq r\},$$

where $\rho_g(h, h') := \mathbb{P}_{D_g}(h(x) \neq h'(x))$.

Definition 7. *Given a group distribution $D_g \in \mathcal{G}$ and a hypothesis class $\mathcal{H}$, let the "disagreement coefficient" of D_g be defined as*

$$\theta_g := \sup_{h \in \mathcal{H}} \sup_{r' \geq 2\nu + \epsilon} \frac{\mathbb{P}_{D_g}(x \in \Delta(B_g(h, r')))}{r'}.$$

We further define the disagreement coefficient over a collection of group distributions $\mathcal{G}$ as

$$\theta_{\mathcal{G}} := \max_{g' \in [G]} \theta_{g'}.$$

Given these definitions, we are now ready to state the main theorem. The consistency comes from what we showed in Lemma 4: as the true error for each $h \in \mathcal{H}_{i+1}$ decreases with each iteration, after enough iterations we will have each $h \in \mathcal{H}_{i+1}$ having ϵ-optimality.

The label complexity bound follows standard ideas in the DBAL literature; see for example [9, 24]. Essentially, what we do is show that at each iteration i, because the true error of any $h \in \mathcal{H}_i$ on the multi-group objective can't be too large, the disagreement of h and h^* on any single group cannot be too large. This leads to a bound on the size of the disagreement region for each g.

Theorem 4. *For all $\epsilon > 0$, $\delta \in (0, 1)$, collections of group distributions $\mathcal{G}$, and hypothesis classes $\mathcal{H}$ with $d < \infty$, with probability $\geq 1 - \delta$, the output $\hat{h}$ of Algorithm 1 satisfies*

$$L_{\mathcal{G}}^{\max}(\hat{h}) \leq L_{\mathcal{G}}^{\max}(h^*) + \epsilon,$$

and its label complexity is bounded by

$$\tilde{O}\left(G\,\theta_{\mathcal{G}}^2\left(\frac{\nu^2}{\epsilon^2} + 1\right)\big(d\log(1/\epsilon) + \log(1/\delta)\big)\log(1/\epsilon) + \frac{G\log(1/\epsilon)\log(1/\delta)}{\epsilon^2}\right).$$

Proof. Lemma 4 says that the number of samples drawn at each iteration is sufficiently large that with probability $\geq 1 - \delta$, for all $i \in [I]$, it holds that for all $h \in \mathcal{H}_{i+1}$, that we have $\left|L_{\mathcal{G}}^{\max}(h) - L_{\mathcal{G}}^{\max}(h^*)\right| \leq 2^{I-i}\epsilon$. Thus, after $I = \lceil \log(1/\epsilon) \rceil$ iterations, the output $\hat{h}$ satisfies the consistency condition.

To see the label complexity, which is the sum of the number of labels we query at each iteration, we note at iteration i, we label no more than

$$1024\left(\frac{m_i}{\epsilon 2^{I-i}}\right)^2\left(2d\log\left(\frac{64}{\epsilon}\right) + \ln\left(\frac{8G\lceil\log(1/\epsilon)\rceil}{\delta}\right)\right) + \frac{128\ln(4G\lceil\log(1/\epsilon)\rceil/\delta)}{\epsilon^2}$$

samples for each group distribution D_g, where $m_i = \max_{g'} \mathbb{P}_{D_{g'}}(x \in \Delta(\mathcal{H}_i))$. The only term here that depends on i is $\frac{m_i}{\epsilon 2^{I-i}}$. By Lemma 4, with probability $\geq 1 - \delta$, it holds for each $i > 1$ that $\left|L_{\mathcal{G}}^{\max}(h) - L_{\mathcal{G}}^{\max}(h^*)\right| \leq 2^{I-i+1}\epsilon$; this holds automatically at $i = 1$ by the setting of $I = \lceil \log(1/\epsilon) \rceil$. Thus, at arbitrary i and for arbitrary $g \in [G]$, we may write

$$\begin{aligned}
\rho_g(h, h^*) &= \mathbb{P}_{D_g}(h(x) \neq h^*(x)) \\
&= \mathbb{P}_{D_g}(h(x) \neq y, h^*(x) = y) + \mathbb{P}_{D_g}(h(x) = y, h^*(x) \neq y) \\
&\leq \mathbb{P}_{D_g}(h(x) \neq y) + \mathbb{P}_{D_g}(h^*(x) \neq y) \\
&= L_{\mathcal{G}}(h \mid g) + L_{\mathcal{G}}(h^* \mid g) \\
&\leq L_{\mathcal{G}}^{\max}(h) + L_{\mathcal{G}}^{\max}(h^*) \\
&= L_{\mathcal{G}}^{\max}(h) - L_{\mathcal{G}}^{\max}(h^*) + L_{\mathcal{G}}^{\max}(h^*) + L_{\mathcal{G}}^{\max}(h^*) \\
&\leq 2^{I-i+1}\epsilon + 2\nu,
\end{aligned}$$

where we recall ν is the noise rate on the multi-group objective. Thus, with probability $\geq 1-\delta$, for each $i \in I$ and $g \in [G]$, it holds that

$$\mathcal{H}_i \subseteq B_g(h^*, 2^{I-i+1}\epsilon + 2\nu).$$

Given this observation, we may then write, for all g, that

$$\mathbb{P}_{D_g}(x \in \Delta(\mathcal{H}_i)) \leq \mathbb{P}_{D_g}(x \in \Delta(B_k(h^*, 2\nu + 2^{I-i+1}\epsilon))),$$

as if there are $h, h' \in \mathcal{H}_i$ that disagree on some x, we have $h, h' \in B_g(h^*, 2\nu + 2^{I-i+1}\epsilon)$, and so h, h' also realize disagreement on x for the larger set of classifiers. Recalling the definition of m_i, this allows us to bound the sum of terms depending on i for each distribution D_g as

$$\begin{aligned}
\sum_{i=1}^{I}\left(\frac{m_i}{\epsilon 2^{I-i}}\right)^2 &\leq \sum_{i=1}^{I}\left(\frac{\max_{g'} \mathbb{P}_{D_g}\left(x \in \Delta(B_{g'}(h^*, 2\nu + 2^{I-i+1}\epsilon))\right)}{2^{I-i}\epsilon}\right)^2 \\
&\leq \sum_{i=1}^{I}\left(\max_{g'} \frac{\mathbb{P}_{D_g}\left(x \in \Delta(B_{g'}(h^*, 2\nu + 2^{I-i+1}\epsilon))\right)}{2\nu + 2^{I-i+1}\epsilon} \cdot \frac{2\nu + 2^{I-i+1}\epsilon}{2^{I-i}\epsilon}\right)^2 \\
&\leq 4\left(\frac{\nu+\epsilon}{\epsilon}\right)^2 \sum_{i=1}^{I}\left(\max_{g'} \frac{\mathbb{P}_{D_g}\left(x \in \Delta(B_{g'}(h^*, 2\nu + 2^{I-i+1}\epsilon))\right)}{2\nu + 2^{I-i+1}\epsilon}\right)^2 \\
&\leq 4\left(\frac{\nu+\epsilon}{\epsilon}\right)^2 \sum_{i=1}^{I}\left(\max_{g'} \sup_{h \in \mathcal{H}} \sup_{r \geq 2\nu+\epsilon} \frac{\mathbb{P}_{D_g}\left(x \in \Delta(B_{k'}(h, r))\right)}{r}\right)^2 \\
&= 4\lceil \log(1/\epsilon) \rceil \left(\frac{\nu+\epsilon}{\epsilon}\right)^2 \left(\max_{g'} \theta_{g'}\right)^2 \\
&= 4\lceil \log(1/\epsilon) \rceil \left(\frac{\nu+\epsilon}{\epsilon}\right)^2 \theta_{\mathcal{G}}^2.
\end{aligned}$$

The label complexity bound then follows by noting the algorithm labels the same amount of samples for all G groups each iteration, and ignoring the factors of $\log(G)$ and $\log(\log(1/\epsilon))$. $\square$

9.2 Group-Realizable Guarantees

Theorem 5. *Suppose Algorithm 2 is run with the active learner $\mathcal{A}_{CAL}$ of [26]. Then for all $\epsilon > 0$, $\delta \in (0,1)$, hypothesis classes $\mathcal{H}$ with $d < \infty$, and collections of group distributions $\mathcal{G}$ that are group realizable with respect to $\mathcal{H}$, with probability $\geq 1-\delta$, the output $\hat{h}$ satisfies*

$$L_{\mathcal{G}}^{\max}(\hat{h}) \leq L_{\mathcal{G}}^{\max}(h^*) + \epsilon,$$

and the number of labels requested is

$$\tilde{O}\Big(dG\theta_{\mathcal{G}} \log(1/\epsilon)\Big).$$

Proof. The label complexity follows directly from the guarantees given in [15]. By a union bound, we with probability $\geq 1-\delta$, have that for all $g \in [G]$, that $\mathcal{A}_{CAL}$ returns $\hat{h}_g$ with the property that

$$L_{\mathcal{G}}(\hat{h}_g \mid g) \leq \epsilon/6.$$

Fix some $g \in [G]$ arbitrarily. Consider a counterfactual training set S_g, unseen by the learner, constructed by labeling each example $x \in S'_g$ via the oracle call $O_g(x)$. Then Vapnik [28] tells us that $m_g := |S'_g|$ is sufficiently large that with probability $\geq 1-\delta/2$, for each $h \in \mathcal{H}$ simultaneously, we have

$$\left|L_{\mathcal{G}}(h \mid g) - L_{S_g}(h)\right| < \epsilon/6.$$

Again by the union bound, this uniform convergence on S_g and the guarantee on the runs of $\mathcal{A}_{CAL}$ both hold for each $g \in [G]$. Conditioned on this high probability event, we can first note that for some arbitrary $h \in \mathcal{H}$,

$$\begin{aligned}
\left|L_{S_g}(h) - L_{\hat{S}_g}(h)\right| &= \left|\frac{1}{m_g}\sum_{i=1}^{m_g} \mathbb{1}[h(x_i) \neq y_i] - \mathbb{1}[h(x_i) \neq \hat{h}_g(x_i)]\right| \\
&\leq \frac{1}{m_g}\sum_{i=1}^{m_g} \left|\mathbb{1}[h(x_i) \neq y_i] - \mathbb{1}[h(x_i) \neq \hat{h}_g(x_i)]\right| \\
&\leq \frac{1}{m_g}\sum_{i=1}^{m_g} \mathbb{1}[y_i \neq \hat{h}_g(x_i)] \\
&= L_{S_g}(\hat{h}_g) \\
&\leq L_{\mathcal{G}}(\hat{h}_g) + \epsilon/6 \\
&\leq \epsilon/6 + \epsilon/6 \\
&= \epsilon/3,
\end{aligned}$$

where the final equality comes from the success of the runs of $\mathcal{A}_{CAL}$. Then for arbitrary h, combining Vapnik's guarantee and the inequality we just showed, we may write:

$$\begin{aligned}
\left|L_{\mathcal{G}}(h \mid g) - L_{\hat{S}_g}(h)\right| &= \left|L_{\mathcal{G}}(h \mid g) - L_{S_g}(h) + L_{S_g}(h) - L_{\hat{S}_g}(h)\right| \\
&\leq \left|L_{\mathcal{G}}(h \mid g) - L_{S_g}(h)\right| + \left|L_{S_g}(h) - L_{\hat{S}_g}(h)\right| \\
&< \epsilon/6 + \epsilon/3 \\
&= \epsilon/2.
\end{aligned}$$

Given this guarantee on the representativeness of the artificially labeled samples on each group g, we have a guarantee for the representativeness over the worst case. For arbitrarily $h \in \mathcal{H}$, we may write

$$\begin{aligned}
\left|L_{\mathcal{G}}^{\max}(h) - \max_{g\in[G]} L_{\hat{S}_g}(h)\right| &= \left|\max_{g\in[G]} L_{\mathcal{G}}(h \mid g) - \max_{g\in[G]} L_{\hat{S}_g}(h)\right| \\
&\leq \max_{g\in[G]} \left|L_{\mathcal{G}}(h \mid g) - L_{\hat{S}_g}(h)\right| \\
&\leq \epsilon/2.
\end{aligned}$$

Thus, by the fact that $\hat{h}$ is the ERM, we have

$$L_{\mathcal{G}}^{\max}(\hat{h}) \leq \max_{g\in[G]} L_{\hat{S}_g}(\hat{h}) + \epsilon/2 \leq \max_{g\in[G]} L_{\hat{S}_g}(h^*) + \epsilon/2 \leq L_{\mathcal{G}}^{\max}(h^*) + \epsilon.$$

□

9.3 Approximation Guarantees

Theorem 6. *Suppose Algorithm 3 is run with the active learner $\mathcal{A}_{DHM}$ of [15]. Then for all $\epsilon > 0$, $\delta \in (0, 1)$, hypothesis classes $\mathcal{H}$ with $d < \infty$, and collections of groups $\mathcal{D}$, with probability $\geq 1 - \delta$, the output $\hat{h}$ satisfies*

$$L_{\mathcal{G}}^{\max}(\hat{h}) \leq L_{\mathcal{G}}^{\max}(h^*) + 2 \cdot \max_{g\in[G]} \nu_g + \epsilon \leq 3 \cdot L_{\mathcal{G}}^{\max}(h^*) + \epsilon,$$

and the number of labels requested is

$$\tilde{O}\left(dG\theta_{\mathcal{G}}\left(\log^2(1/\epsilon) + \frac{\nu^2}{\epsilon^2}\right)\right).$$

Proof. The proof is almost identical to that of Theorem 2. The label complexity bound follows directly from [10]. Similar to before, we have that for all $g \in [G]$, $\mathcal{A}_{DHM}$ returns $\hat{h}_g$ with the property that

$$L_{\mathcal{G}}(\hat{h}_g \mid g) \leq L_{\mathcal{G}}(h_g^* \mid g) + \epsilon/6.$$

Fix some $g \in [G]$ arbitrarily. On a counterfactual training set S_g, unseen by the learner, constructed by labeling each example $x \in S'_g$ via the oracle call $O_g(x)$, it holds that $m_g := |S'_g|$ is sufficiently large that with probability $\geq 1 - \delta/2$, for each $h \in \mathcal{H}$ simultaneously, we have

$$\left|L_{\mathcal{G}}(h \mid g) - L_{S_g}(h)\right| < \epsilon/6.$$

By the union bound, this uniform convergence and the guarantee on the runs of $\mathcal{A}_{DHM}$ both hold. Thus, we can first note that for some arbitrary $h \in \mathcal{H}$,

$$\begin{aligned}
\left|L_{S_g}(h) - L_{\hat{S}_g}(h)\right| &= \left|\frac{1}{m_g}\sum_{i=1}^{m_g} \mathbb{1}[h(x_i) \neq y_i] - \mathbb{1}[h(x_i) \neq \hat{h}_g(x_i)]\right| \\
&\leq \frac{1}{m_g}\sum_{i=1}^{m_g} \left|\mathbb{1}[h(x_i) \neq y_i] - \mathbb{1}[h(x_i) \neq \hat{h}_g(x_i)]\right| \\
&\leq \frac{1}{m_g}\sum_{i=1}^{m_g} \mathbb{1}[y_i \neq \hat{h}_g(x_i)] \\
&= L_{S_g}(\hat{h}_g) \\
&\leq L_{\mathcal{G}}(\hat{h}_g \mid g) + \epsilon/6 \\
&\leq L_{\mathcal{G}}(h^*_g \mid g) + \epsilon/3 \\
&= \nu_g + \epsilon/3.
\end{aligned}$$

where the second to last inequality comes from uniform convergence over S_G, and the final equality comes from the correctness guarantee of $\mathcal{A}_{DHM}$. Then for arbitrary h, combining Vapnik's guarantee and the inequality we just showed, we may write:

$$\begin{aligned}
\left|L_{\mathcal{G}}(h \mid g) - L_{\hat{S}_g}(h)\right| &= \left|L_{\mathcal{G}}(h \mid g) - L_{S_g}(h) + L_{S_g}(h) - L_{\hat{S}_g}(h)\right| \\
&\leq \left|L_{\mathcal{G}}(h \mid g) - L_{S_g}(h)\right| + \left|L_{S_g}(h) - L_{\hat{S}_g}(h)\right| \\
&< \epsilon/6 + \nu_g + \epsilon/3 \\
&= \nu_g + \epsilon/2.
\end{aligned}$$

Then, as above, we have, for arbitrarily $h \in \mathcal{H}$,

$$\left|L^{\max}_{\mathcal{G}}(h) - \max_{g\in[G]} L_{\hat{S}_g}(h)\right| \leq \max_{g\in[G]} \left|L_{\mathcal{G}}(h \mid g) - L_{\hat{S}_g}(h)\right| \leq \max_{g\in[G]} \nu_g + \epsilon/2 \leq \nu + \epsilon/2,$$

where the the final inequality comes from the fact that if any hypothesis has less than ν_g error on all groups, it would be optimal on group g. Thus, by the fact that $\hat{h}$ is the ERM, we have

$$L^{\max}_{\mathcal{G}}(\hat{h}) \leq \max_{g\in[G]} L_{\hat{S}_g}(\hat{h}) + \nu_g + \epsilon/2 \leq \max_{g\in[G]} L_{\hat{S}_g}(h^*) + \nu_g + \epsilon/2 \leq L^{\max}_{\mathcal{G}}(h^*) + 2\nu + \epsilon \leq 3 \cdot L^{\max}_{\mathcal{G}}(h^*) + \epsilon$$

$\square$