Agnostic Multi-Group Active Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Inspired by the problem of improving classification accuracy on rare or hard subsets of a population, there has been recent interest in alternative models of learning where the goal is to generalize to a collection of distributions, each representing a "group". We consider a variant of this problem from the perspective of active learning, where the learner is endowed with the power to decide which examples are labeled from each distribution in the collection, and the goal is to minimize the number of label queries while maintaining PAC-learning guarantees. We demonstrate an active learning algorithm for an agnostic formulation of this problem, which given a collection of distributions of size G and hypothesis class $\mathcal{H}$ with VC-dimension d, outputs an ϵ-suboptimal hypothesis using $\tilde{O}\left(Gd\log^2(1/\epsilon) + G/\epsilon^2\right)$ label queries when disagreement coefficients are bounded independently of ϵ. When $G < o(\log^2(1/\epsilon)/\epsilon^2)$, this guarantee is of strictly lower order than sample complexity lower bounds for a learner that may decide how many samples it wants from each distribution in the collection during training. We also consider the special case of the problem where each distribution in the collection is individually realizable with respect to $\mathcal{H}$, and demonstrate $\tilde{O}\left(Gd\log(1/\epsilon)\right)$ label queries are sufficient for learning in this case. We further give an approximation result for the full agnostic case inspired by the group realizable strategy.

1 Introduction

There has been growing interest in learning problems where the goal is to choose a classifier that performs well when faced with multiple subpopulations or "groups", each with their own distribution [1, 2, 3, 4, 5, 6, 7]. In many cases, the motivation comes from a perspective of fairness, where a typical requirement is that we classify with similar accuracy across groups [7, 6]. In other cases, the motivation may simply be to train more reliable classifiers. For example, it has been observed that cancer detection models with good overall accuracy often suffer from poor ability to detect rare subtypes of cancer that are not well-represented or identified in training [8].

In this work, we consider the following formulation of the "multi-group" problem. The learner is given a collection of distributions $\{D_g\}_{g=1}^{G}$, each corresponding to a group, a hypothesis class $\mathcal{H}$, and wants to pick a classifier that approximately minimizes the maximum classification error over group distributions. We consider this problem from an active learning perspective, where the learner has the power to choose which examples from each group it wants to label during training. In a standard extension of the active learning literature, we set out to design schemes for choosing which examples from each group should be labeled, where the goal is to minimize the number of label queries while retaining PAC-learning guarantees.

A major challenge in designing active learning strategies for the multi-group problem is that disagreement based active learning (DBAL) - the most successful algorithmic paradigm for agnostic active learning - fails to admit naive application in the multi-group setting. In DBAL, a standard idea is

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

that at any time during training, an active learner may safely abstain from requesting labels outside a region of space called the "disagreement region", a subset of instance space where empirically well-performing hypotheses disagree about how new examples should be labeled. When the learner need only consider a single distribution, error differences between classifiers are specified entirely through their performance on the part of space on which they disagree, i.e. the disagreement region. However, when multiple group distributions must be considered, the absolute errors of classifiers on each group must be estimated to compare performance of two classifiers, and this property no longer holds. We resolve this via the observation that while we cannot spend all our labeling budget in the disagreement region, we can exploit the agreement in its complement to cheaply estimate absolute errors of classifiers on each group. This leads to our performance gains in the full agnostic setting.

In this setting, we demonstrate a consistent active learning algorithm which relies on our modification of standard ideas in DBAL referenced above. We analyze the number of label queries made by this scheme in terms of a standard complexity measure in the active learning literature called the "disagreement coefficient" [9, 10], and show that $\tilde{O}\left(G\log(|\mathcal{H}|)\log^2(1/\epsilon) + G/\epsilon^2\right)$ labels queries sufficient for our specification of multi-group learning, when disagreement coefficients of each of the group distributions D_g are bounded independently of ϵ. We note that when $G < o(\log(1/\epsilon/\epsilon^2))$, our scheme uses fewer labels than required by lower bounds for a variant of the problem where the learner may only specify how many examples from each group to draw during training. We also show that all dependence on $1/\epsilon^2$ in the label complexity can be replaced with $\log(1/\epsilon)$ when each distribution is individually realizable, and give approximation results for the general agnostic case of multi-group learning with $\log(1/\epsilon)$ dependence on ϵ.

2 Related Work

2.1 Multi-Group Learning

The majority of the empirical work on multi-group learning has been through the lens of "Group-Distributionally Robust Optimization" (G-DRO) [11, 12, 13]. In the former case, the goal is to choose a classifier that minimizes the maximal risk against an unknown mixture over a collection of distributions $\{D_g\}_{g=1}^G$ representing groups. One assumes a completely passive sampling setting – all data is given to the learner at the beginning of training, and the learner has no ability to draw extra, fine-grained samples. The strategy is usually empirical risk minimization (ERM) or some regularized ERM variant over the max loss - for a set of classifiers parameterized by $\phi \in \Phi$, and letting S_g denote the set of examples in the training set coming from D_g, one performs

$$\min_{\phi\in\Phi}\max_{g\in[G]}\frac{1}{|S_g|}\sum_{(x_i,y_i)\in S_g} l(f_\phi(x_i), y_i)$$

It is important to note that the learner knows the group identity of each sample in the training set, but is not provided with group information, precluding the possibility of training a separate classifiers for each group.

"Multi-group PAC learning" consider the multi-group problem under the passive sampling assumption from a more classical learning-theoretic perspective [7, 6]. Here, one assumes there is a single distribution D from which one is given samples, but also a collection of subsets of instance space $\mathcal{G}$ over which one wants to learn conditional distributions. Given a hypothesis class $\mathcal{H}$, the learner tries to improperly learn a classifier f that competes with the optimal hypothesis on each conditional distribution specified by a group g in the collection - formally, one requires that for a given error tolerance ϵ, f has the property

$$\forall g \in \mathcal{G},\ \ \mathbb{P}_{(x,y)\sim D}(f(x) \neq y | x \in g) \leq \inf_{h\in\mathcal{H}} \mathbb{P}_{(x,y)\sim D}(h(x) \neq y | x \in g) + \epsilon$$

with high probability. An interesting wrinkle in this literature is that the group identity of samples is available at both training and test times. It has been shown that a sample complexity of $\tilde{O}\left(\log(|\mathcal{G}||\mathcal{H}|)/\gamma\epsilon^2\right)$ is sufficient for learning in this model, where γ is the minimal mass of a group g under D [6].

"Collaborative learning" studies the multi-group problem under an alternative sampling model [1, 2, 3]. Here we are given a collection of distributions $\{D_g\}_{g=1}^G$, each corresponding to a group. Given some

hypothesis class $\mathcal{H}$, the goal is to learn a classifier f, possibly improperly, that is evaluated against its worst-case loss over $D_1, \ldots, D_G$; formally, we would like f to satisfy

$$\max_{g \in [G]} \mathbb{P}_{(x,y)\sim D_g}(f(x) \neq y) \leq \inf_{h \in \mathcal{H}} \max_{g \in [G]} \mathbb{P}_{(x,y)\sim D_g}(h(x) \neq y) + \epsilon.$$

In contrast with multi-group PAC learning, the learner may decide how many samples from each D_g it wants to collect during training, and group identity is hidden on test examples. This models the case where a learner may want to collect more data from a particularly difficult group of instances, such as a rare or hard-to-diagnose type of cancer. Recently, it was shown for finite hypothesis classes that $\tilde{\Theta}(\log(|\mathcal{H}|)/\epsilon^2 + G/\epsilon^2)$ total samples over all groups are necessary and sufficient to learn in this model [3].

Our work builds further on collaborative learning, and endows the learner with the ability decide which samples from each group distribution D_g should be labeled. This is the standard framework of active learning, applied to the multi-group setting.

2.2 Active Learning

Active learning concerns itself with the development of learning algorithms for training classifiers that have power over which training examples should be labeled [14, 15]. Much work in the field has focused on uncovering settings in which algorithmic approaches lower the amounts of labels required for PAC-style learning guarantees beyond lower bounds that apply when data is collected i.i.d. from the target distribution [16, 17]. In the agnostic, 0-1 loss setting, "lower" ideally generally means reducing the dependence on ϵ in label complexities from a multiplicative factor of $1/\epsilon^2$ to one of $\text{polylog}(1/\epsilon)$.

The vast majority of the work on active learning has been done in the 0-1 loss setting, where accuracy over a single, fixed test distribution is the measure of performance. This setting is fairly well-understood, at least in the sense that a significant body of work has arisen demonstrating the power of active learning to reduce so-called "label complexities" - the number of label queries made by the active learner [18, 9, 10, 19, 20, 21]. It has been significantly harder to push the design of active learning algorithms past the regime of accuracy on a fixed target. While some work has attempted to generalize classical ideas of active learning to different losses [22], these are heavily outnumbered in the literature, and are accompanied by some negative results describing the inability of active learning to improve over passive learning in certain settings outside the 0-1 loss setting [23].

The efficacy of active learning is known to depend on certain "niceness" conditions of the data generating distribution. In particular, in the agnostic case, the reduction of the dependence on the label complexity from $1/\epsilon^2$ to $\text{polylog}(1/\epsilon)$ requires the accuracy of the optimal hypothesis in the hypothesis class to be high, and a parameter called the "disagreement coefficient" - describing the easiness of eliminating significantly suboptimal hypotheses from contention - to be bounded [15]. In the worst case, the power to selectively label examples from the target provides no gain in label complexity over i.i.d. sampling from the target [24, 22].

3 Preliminaries

3.1 Learning Problem

We study a binary classification setting where examples fall in some instance space $\mathcal{X}$, and labels lie in $\mathcal{Y} := \{-1, 1\}$. We suppose we are given some pre-specified, finite collection of distributions $\mathcal{G} = \{D_g\}_{g=1}^{G}$ over $\mathcal{X} \times \mathcal{Y}$ corresponding to groups; for a given group index g, let μ_g denote the marginal measure over instance space of D_g. Given a hypothesis class $\mathcal{H}$ of classifiers with VC-dimension d, the goal of the leaner is to pick some $h \in \mathcal{H}$ from finite data that performs well across all the distributions in $\mathcal{D}$ in the worst case. Let

$$L_{\mathcal{D}}(h \mid g) := \mathbb{P}_{(x,y)\sim D_g}(h(x) \neq y)$$

be the error of a hypothesis h on group g. Formally speaking, the learner would like to choose a classifier approximately obtaining

$$\inf_{h \in \mathcal{H}} \max_{g \in [G]} L_{\mathcal{G}}(h \mid g)$$

using finite data. We often use $L_{\mathcal{D}}^{\max}(h)$ as shorthand for $\max_{g \in [G]} L_{\mathcal{G}}(h \mid g)$. We use $\nu := \inf_{h \in \mathcal{H}} L_{\mathcal{D}}^{\max}(h)$ to denote the "noise rate" of $\mathcal{H}$ on the multi-distribution objective. We assume for simplicity that there is some $h^* \in \mathcal{H}$ attaining ν. The use of the term "agnostic" throughout reflects the fact that we make no assumption that $\nu = 0$ in our algorithm design or analysis.

Because we consider this problem from an active learning perspective, it is important that we search for learning strategies that are "consistent" in the sense that as the number of number of labels requested approaches infinity, the learner outputs the true optimal hypothesis. Consistency is an important property that can easy fail for agnostic active learning strategies [15].

3.2 Active Learning Model

We consider a standard active learning model specified as follows. Let $supp(D_g)$ denote the support of μ_g. The active learner has access to two sampling oracles for each distribution specified by D_g. The first is $U_g(\cdot)$, which given a set $S \subseteq \mathcal{X}$ measurable with respect to μ_g, returns an unlabeled sample from μ_g conditioned on S; if $S \cap supp(D_g) = \emptyset$, $U_g(S)$ returns "None". The second is $O_g(\cdot)$, which given a point in $supp(D_g)$, returns a sample from the conditional distribution over labels specified by x and g. More formally, querying $U_g(S)$ for $S \cap supp(D_g) \neq \emptyset$ is equivalent to drawing i.i.d. samples according to μ_g (independent of previous randomness), and returning the first example that falls in $S \cap supp(D_g)$; querying the oracle $O_g(x)$ for $x \in supp(D_g)$ is equivalent to receiving a sample from a Bernoulli random variable with parameter $\mathbb{P}_{(x,y)\sim D_g}(Y = 1 | X = x)$.

As is standard in active learning, the active learner is assumed to have functionally unlimited access to queries from $U_g(\cdot)$. On the other hand, queries to oracles $O_g(\cdot)$ are precious: the "label complexity" of a strategy executed by the active learner is the sum of queries to oracles $O_g(\cdot)$ over all g, and is to be minimized given a desired generalization error guarantee.

3.3 Measurability

To avoid unnecessarily complicated discussion of measure and σ-field, we assume throughout that for each $\mathcal{H}' \subseteq \mathcal{H}$, we have that $\Delta(\mathcal{H}') \cap supp(D_g)$ is measurable under μ_g for each g, where

$$\Delta(\mathcal{H}') := \{x \in \mathcal{X} : \exists h, h' \in \mathcal{H}' \; s.t. \; h(x) \neq h'(x)\}$$

is the "disagreement region" of $\mathcal{H}'$. When a single probability space generates the data, measurability of this region follows from measurability of the functions in $\mathcal{H}$ [9].

4 Active Learning: From a Single Distribution to Multi-Group Learning

We base our full agnostic algorithm on DBAL ideas. In this section, we give some background on classical DBAL on a single distribution, and discuss in more detail the challenges facing DBAL in the case of multi-group learning.

4.1 Background on Disagreement-Based Active Learning

Arguably the most successful idea in agnostic active learning for accuracy over a single distribution has been so-called "disagreement-based active learning" [18, 10, 16]. The essential idea in this school of algorithms is that one can learn the relative accuracy of two classifiers h and h' by only requesting labels for examples in the part of instance space on which they disagree about how examples should be labeled. More generally, given a set of classifiers $\mathcal{H}' \subseteq \mathcal{H}$, one can consider the "disagreement region" of $\mathcal{H}'$

$$\Delta(\mathcal{H}') := \{x \in \mathcal{X} : \exists h, h' \in \mathcal{H}' \; s.t. \; h(x) \neq h'(x)\}.$$

As alluded to above, the difference in accuracy of classifiers $h, h' \in \mathcal{H}'$ is specified entirely through this inherently label-independent notion. Fix a single distribution D, we have

$$\begin{aligned}&\mathbb{P}_{(x,y)\sim D}\left(h(x) \neq y \mid x \in \Delta(\mathcal{H}')\right) - \mathbb{P}_{(x,y)\sim D}\left(h'(x) \neq y \mid x \in \Delta(\mathcal{H}')\right) \\ &= \frac{\mathbb{P}_{(x,y)\sim D}\left(h(x) \neq y\right) - \mathbb{P}_{(x,y)\sim D}\left(h'(x) \neq y\right)}{\mathbb{P}_{(x,y)\sim D}\left(x \in \Delta(\mathcal{H}')\right)},\end{aligned}$$

as by virtue of their agreement, h, h' have the same conditional loss on $\Delta(\mathcal{H}')^c$. Inspired by this observation, the fundamental idea is to label examples in $\Delta(\mathcal{H}')$, and ignore those outside of it, thus

saving on labeling examples that are not informative of the relative classification ability members of $\mathcal{H}'$. Ideally, certain classifiers quickly reveal themselves to be so much worse than the current ERM hypothesis, that by standard concentration bounds, they can be inferred to be more than ϵ-suboptimal with high probability. Elimination of these classifiers shrinks the disagreement region, and allowing the labeling to safely become further fine-grained.

4.2 Breakdown of Standard Disagreement Methods

In the multi-target setting, the utility of examining the disagreement of two classifiers is weakened. The problem is as follows: although the classifiers in $\mathcal{H}'$ still agree in the complement of $\Delta(\mathcal{H}')$, this is no longer enough infer differences in the worst case error over groups $L_{\mathcal{D}}^{\max}$. To make claims about the differences in worst case error, one can no longer naively check performance in $\Delta(\mathcal{H}')$, because differences on $\Delta(\mathcal{H}')$ are not generally representative of absolute errors over target distributions. The following simple example makes this concrete.

Example 1. Consider the task of determining which of two classifiers h and h' has lower error in the worst case over distributions D_1 and D_2 with marginal supports $S_1 \subseteq \mathcal{X}$ and $S_2 \subseteq \mathcal{X}$. Let their disagreement region be denoted by $\Delta = \{x \in \mathcal{X} : h(x) \neq h'(x)\}$, and let $risk(f, i, S)$ denote the conditional risk of classifier f on $S_i \cap S$ under D_i. Suppose we only know their conditional risks on $\Delta \cap S_1$ and $\Delta \cap S_2$ under D_1 and D_2, respectively. We see for h that

$$risk(h, i, S) = \begin{cases} 1/4 & i = 1, S = \Delta \cap S_1 \\ 1/3 & i = 2, S = \Delta \cap S_2 \\ ? & i = 1, S = \Delta^c \cap S_1 \\ ? & i = 2, S = \Delta^c \cap S_2 \end{cases}$$

and for h' that

$$risk(h', i, S) = \begin{cases} 34/100 & i = 1, S = \Delta \cap S_1 \\ 0 & i = 2, S = \Delta \cap S_2 \\ ? & i = 1, S = \Delta^c \cap S_1 \\ ? & i = 2, S = \Delta^c \cap S_2 \end{cases}.$$

Consider ignoring risks in Δ^c, and using as a surrogate for the multi-group objective

$$\max_{i \in \{1,2\}} risk(h, i, S_i \cap \Delta).$$

In this case, we would chose h has the better of the two hypotheses. However, suppose further that $\Delta \cap S_1$ and $\Delta \cap S_2$ have mass $1/2$ under both D_1 and D_2, respectively, and that $risk(h, 1, \Delta \cap S_1) = risk(h', 1, \Delta \cap S_1) = 1/1000$, and that $risk(h, 2, \Delta \cap S_2) = risk(h', 2, \Delta \cap S_2) = 1/2$. Then one can compute that h' has a lower worst case error over groups D_1 and D_2.

5 General Agnostic Multi-Group Learning

5.1 An Agnostic Algorithm

The basic idea in Algorithm 1 is similar to classical active learning approaches for a single distribution. We start with the full hypothesis class $\mathcal{H}$, and look to iteratively eliminate hypotheses from contention as we learn about how to classify on each groups through targeted labeling. To do this, we construct empirical estimates for the worst case losses over groups, and then iteratively eliminate hypotheses with empirical losses so large that standard concentration arguments imply they cannot be ϵ-suboptimal with high probability.

In order to achieve labeling savings, our algorithm spends a significant amount of its labeling budget in the disagreement region. However, as noted above, we cannot fully ignore labeling in the complement of the disagreement region in the multi-group setting. It is a key observation of ours that we can still exploit the agreement of the remaining classifiers to estimate the absolute error of all hypotheses in contention on a given group with just $O(\log(1/\delta)/\epsilon^2)$ more labels than would otherwise be necessary. To do this, we construct a two-part estimate for the loss of a hypothesis on a given group. Denote the set of hypotheses still in contention at iteration i is $\mathcal{H}_i$. Let $R_i = \Delta(\mathcal{H}_i)$ and $S_{R_i,g}$ be a labeled

sample from $U(supp(D_g) \cap R_i)$ and $S_{R_i^c,g}$ be a labeled sample from $U(supp(D_g) \cap R_i^c)$. We can how estimate the loss for some $h \in \mathcal{H}_i$ on group g via

$$L_{S;R_i}(h \mid g) := \mu_g(R_i) \cdot L_{S_{R_i,g}}(h) + \mu_k(R_i^c) \cdot L_{S_{R_i^c,g}}(h_{\mathcal{H}_i}),$$

where $L_{\mathcal{S}}(h) := \frac{1}{S} \sum_{(x,y)\in\mathcal{S}} \mathbb{1}[h(x) \neq y]$ is a standard empirical loss estimate, and $h_{\mathcal{H}_i}$ is an *arbitrarily* chosen hypothesis from $\mathcal{H}_i$ that is used in the loss estimate of every $h \in \mathcal{H}_i$. This leads to an unbiased estimator given that every $h \in \mathcal{H}_i$ labels the sample from this part of space in exactly the same way.

Add in that when empty set, we set empirical estimate to an arbitrary constant

The utility of this estimator is that by choosing an arbitrary representative $h_{\mathcal{H}_i}$, we can estimate the loss of all hypothesis still in contention to precision $O(\epsilon)$ on $R_i^c \cap supp(D_g)$ with $O(\log(1/\delta)/\epsilon^2)$ samples, removing the usual dependence of the VC-dimension. On the other hand, as the disagreement region shrinks, $\mu_g(R_i)$ shrinks as well, so while we will still need to invoke uniform convergence to get reliable loss estimates in $R_i \cap supp(D_g)$, the precision to which we need to estimate losses in this part of space decreases with every iteration, and eventually the overall dependence on the VC-dimension is diminished. This later observation is the standard source of gains in DBAL [18, 9, 26].

At each iteration, we use this two-part estimator on each group to construct unbiased loss estimates for the worst case over groups via

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

We draw enough samples at each iteration i such that we essentially learn the multi-group problem to precision $2^{\lceil \log(1/\epsilon)\rceil - i}\epsilon$.

We note that Algorithm 1 assumes access to the underlying group marginals measures μ_g. This is common in the active learning literature [18, 20]. Probabilities of events in instance space can be estimated to arbitrary accuracy using only unlabeled data, so this assumption is not dangerous to our goal of lowering label complexities. We also note that while Algorithm 1 is not "executable" as stated for infinite hypothesis classes, ϵ-covers of near-optimal size can be constructed with high probability using a polynomial number of purely unlabeled examples [9].

eventually, i would just say that when you make an oracle call to Ug and the measure is 0, you get back empty set, and go back to non-cases algorithm definition, but this is nice and explict for now. if you do that, be sure to remind the reader of what is going on so the two-part estimator collapse is clear in the body

5.2 Guarantees

The scheme given in Algorithm 1 is consistent. Given that we essentially learn to precision $2^{\lceil \log(1/\epsilon)\rceil - i}\epsilon$ at iteration i, after $\lceil \log(1/\epsilon)\rceil$ iterations, the ERM hypothesis on $L^{\max}_{\mathcal{S};R_i}(\cdot)$ is then ϵ-suboptimal with high probability.

We can bound the label complexity of the algorithm using standard techniques from disagreement-based active learning. A ubiquitous quantity in the analysis of disagreement-based schemes is that of the "disagreement coefficient" [9, 25]. The general idea is that the disagreement coefficient bounds the rate of decrease in r of the measure of the the disagreement region of a ball of radius r around h^* in the pseudo-metric $\rho_g(h,h') := \mathbb{P}_{(x,y)\sim D_g}(h(x) \neq h'(x))$. Precisely, we use the following definition of the disagreement coefficient in our analysis [10, 26]. Given a group D_g, the disagreement coefficient on g is

$$\theta_g := \sup_{h\in\mathcal{H}} \sup_{r'\geq r} \frac{\mu_g\left(\Delta(B_g(h,r'))\right)}{r'},$$

make clear the independence between everything (like you did on the last paper)

Kamalika says eps-suboptimal -> eps-optimal

where $B_g(h,r') := \{h' \in \mathcal{H} : \rho_g(h,h') \leq r'\}$ is a ball of radius r' about h in pseudo-metric ρ_g. We further notate the maximum disagreement coefficient over the groups $\mathcal{G}$ as $\theta_{\mathcal{G}} := \max_g \theta_g$. The disagreement coefficient θ_g is trivially bounded by $1/(2\nu_g + \epsilon)$, but can be bounded independently of ϵ in many cases [10, 25]. For example, the disagreement coefficient when $\mathcal{H}$ is linear separators in d dimensions is $\Theta(\sqrt{d})$ when the underlying distribution is the uniform distribution over the l_2 ball. We state the following guarantee for Algorithm 1 in terms of the disagreement coefficient.

script G instead of script D is nicer I think and fits "groups" idea better

Check this. what's the orginal citation? Also add in other examples, like Tsybakov noise etc

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

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

and its label complexity is bounded by

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

Algorithm 1 General Agnostic Algorithm

procedure MULTI_GROUP_AGNOSTIC($\mathcal{H}, \epsilon, \delta, \{U_g(\cdot)\}_{g=1}^G, \{O_g(\cdot)\}_{g=1}^G$)
  $\mathcal{H}_1 \leftarrow \mathcal{H}$, $I \leftarrow \lceil \log_2(1/\epsilon) \rceil$
  for $i \in [I]$ **do**
    $R_i \leftarrow \Delta(\mathcal{H}_i)$
    $m_i \leftarrow \max_{g' \in [G]} \mu_{g'}(\Delta(\mathcal{H}_i))$
    for $g \in [G]$ **do**
      $\mathcal{S}'_{R_i,g} \leftarrow 2048 \left(\frac{m_i}{\epsilon 2^{I-i}}\right)^2 \left(2d \log(\frac{128}{\epsilon}) + \ln(\frac{8G\lceil \log(1/\epsilon) \rceil}{\delta})\right)$ i.i.d. samples from $U_g(R_i \cap supp(D_g))$
      $\mathcal{S}'_{R_i^c,g} \leftarrow \frac{128 \ln(4/\delta)}{(\epsilon 2^{I-i})^2}$ i.i.d. samples from $U_g(R_i^c \cap supp(D_g))$
      if "None" $\in \mathcal{S}'_{R_i,g}$ **then** ▷ $R_i \cap supp(D_g) = \emptyset$ in this case
        $\mathcal{S}_{R_i,g} \leftarrow \emptyset$
      else
        $\mathcal{S}_{R_i,g} \leftarrow \{(x, O_g(x)) : x \in \mathcal{S}'_{R_i,g}\}$
      end if
      if "None" $\in \mathcal{S}'_{R_i^c,g}$ **then**
        $\mathcal{S}_{R_i^c,g} \leftarrow \emptyset$
      else
        $\mathcal{S}_{R_i^c,g} \leftarrow \{(x, O_g(x)) : x \in \mathcal{S}'_{R_i^c,g}\}$
      end if
    end for
    $\hat{h}_i = \arg\min_{h \in \mathcal{H}_i} L^{\max}_{\mathcal{S};R_i}(h)$
    $\mathcal{H}_{i+1} \leftarrow \left\{h \in \mathcal{H}_i : L^{\max}_{\mathcal{S};R_i}(h) \leq L^{\max}_{\mathcal{S};R_i}(\hat{h}_i) + 2^{I-i}\epsilon/4\right\}$
  end for
  return $\hat{h} = \arg\min_{h \in \mathcal{H}_{I+1}} L^{\max}_{S;R_{I+1}}(h)$
end procedure

The $\tilde{O}$ notation hides factors of $\log(\log(1/\epsilon))$ and $\log(G)$; we leave all proofs for the Appendix.

The implication of Theorem 1 is that when $\theta_{\mathcal{G}}$ can be bounded independently of ϵ, the gain of Algorithm 1 is that the dependence on the standard interaction of the VC-dimension d and $1/\epsilon^2$ is removed, and replaced with $Gd\log^2(1/\epsilon)$.

5.3 Comparison to Collaborative Learning Lower Bounds

The active learning literature has traditionally focused on discovering settings in which active algorithms have lower label complexities than "passive" lower bounds on sample complexity. We compare our label complexity guarantees to the lower bounds for collaborative learning, a strictly stronger comparison than comparing to passive learning.

In collaborative learning [1, 2, 3], the objective is the same as ours - we wish to output a classifier that has low error in the worst case over groups. The only difference is the learner may only specify how many samples from each group it wants during training, and so, does not have the power to selective label examples within a group. [3] show that for finite hypothesis classes $\mathcal{H}$,

$$\Omega\left(\frac{\log(|\mathcal{H}|)}{\epsilon^2} + \frac{G \cdot \log(\min(|\mathcal{H}|, G)/\delta)}{\epsilon^2}\right)$$

total samples over all groups are necessary with this group-conditional sampling power to learn to error ϵ with probability $\geq 1 - \delta$. Thus, Algorithm 1 uses asymptotically less label queries whenever $G < o(\log^2(1/\epsilon)/\epsilon^2)$ and $\theta_{\mathcal{D}}$ is bounded independently of ϵ.

Removing the multiplicative factor G in the first term of our label complexity bounds may be possible with similar ideas, but will require a more refined algorithmic approach - it is not due to slack in our

Algorithm 2 Group-Realizable Algorithm

procedure GROUP_REALIZABLE($\mathcal{H}, \epsilon, \delta$, active learner $\mathcal{A}$, $\{U_g(\cdot)\}_{g=1}^G, \{O_g(\cdot)\}_{g=1}^G$)
  for $g \in [G]$ **do**
    $\hat{h}_g \leftarrow \mathcal{A}(\mathcal{H}, \epsilon/6, \delta/2G, U_g(\mathcal{X}), O_g)$
    $S'_g \leftarrow 144/\epsilon^2 \left(2d\ln(24/\epsilon) + \ln(48G/\delta)\right)$ samples from oracle $U_g(\mathcal{X})$
    $\hat{S}_g \leftarrow \left\{\left(x, \hat{h}_g(x)\right) : x \in S'_g\right\}$
  end for
  return $\hat{h} = \arg\min_{h\in\mathcal{H}} \max_{g\in[G]} \frac{1}{|\hat{S}_g|} \sum_{(x,\hat{y})\in\hat{S}_g} \mathbb{1}\left[h(x) \neq \hat{y}\right]$
end procedure

analysis. This would essentially mean creating an algorithm that beats collaborative learning lower bounds for arbitrarily large collections of groups, which is an interesting topic for future study.

6 Group-Realizable Learning

A special case of the learning problem where more extreme active learning can be readily seen, comes when the hypothesis class $\mathcal{H}$ achieves zero noise rate on each group D_g. This setting has been considered in the passive "multi-group learning" literature [6]. Formally speaking, in the group realizable setting, the following condition holds:

$$\forall g \in [G], \exists h^*_g \in \mathcal{H} \; s.t. \; L_{\mathcal{G}}(h^*_g \mid g) = 0,$$

i.e. for all groups in the collection $\mathcal{G}$, there is some hypothesis achieving 0 error on that group. Note that this differs from the fully realizable setting where there is some $h^* \in \mathcal{H}$ with $L^{\max}_{\mathcal{G}}(h^*) = 0$. While fully realizable implies group realizable, the converse is not true. Thus, group-realizability represents an intermediate regime between the realizable setting and the full agnostic settings.

6.1 Algorithm

In the group-realizable case, it is possible to show a reduction of the problem of active learning over hypothesis classes with respect to a single distribution.

The algorithm examines each group in sequence. For each D_g, it calls as a subroutine an active learner that is guaranteed to find an order ϵ-optimal hypothesis $\hat{h}_g \in \mathcal{H}$ with high probability over it's queries to $U_g(\cdot)$ and $O_g(\cdot)$. It then gathers new unlabeled samples from each D_g, and instead of requesting labels from $O_g(\cdot)$, labels each unlabeled point stemming from $U_g(\cdot)$ with the learned classifier $\hat{h}_g$. The final step is to do an empirical risk minimization on these artificially labeled samples with respect to the multi-group objective. See Algorithm 2 for a formal specification of the strategy.

6.2 Guarantees

The strategy given in Algorithm 2 leads to a consistent active learning scheme, provided the active learners called as subroutines have standard guarantees that can be inherited.

Theorem gives a finite sample guarantee for Algorithm 2 . The proof follows from an argument similar to one used in [**?**] - the subroutine calls return hypotheses with near 0 error on each group, and so the artificially labeled training set used in the ERM step looks nearly identical to a counterfactual training set for the ERM step constructed by querying labels $O_g(x)$ for each unlabeled x. We present Theorem assuming access to a classical, realizable active learner due to [27], noting that tighter label complexity bounds are possible in certain scenarios using more advanced single-distribution active learning schemes.

Theorem 2. *Fix an arbitrary $\mathcal{H}$ with $d < \infty$, a collection $\mathcal{G}$ that is distribution-wise realizable with respect to $\mathcal{H}$, and suppose $\mathcal{A}$ is that of [27]. Then for all settings of ϵ, δ, with probability $\geq 1 - \delta$, the output $\hat{h}$ of Algorithm 2 satisfies*

$$L^{\max}_{\mathcal{G}}(\hat{h}) \leq \inf_{h\in\mathcal{H}} L^{\max}_{\mathcal{G}}(h) + \epsilon,$$

and its label complexity is

$$\tilde{O}\left(Gd\,\theta_{\mathcal{G}}\log(1/\epsilon)\right).$$

The best realizable active learners are often much more "aggressive" in their querying strategy than their agnostic counterparts, and thus have much lower label complexities. Exploiting this fact and the reduction of Algorithm 2, we emphasize that the label complexity guarantee of Theorem 3 is much stronger than what one would achieve from agnostically learning over $\mathcal{H}$ using Algorithm **??**. When disagreement coefficients across the collection of targets are bounded independently of ϵ, the dependence on $1/\epsilon^2$ is replaced by $\log(1/\epsilon)$. We note that the disagreement coefficient $\theta_{\mathcal{G}}$ is trivially bounded above by $1/\epsilon$, so in the worst case, this label complexity guarantee is bounded above by $\tilde{O}\left(Kd/\epsilon\right)$, removing the standard dependence of $1/\epsilon^2$ from passive learning.

7 The Gap Between Group Realizable and Full Agnostic

7.1 Inconsistency of the Reduction in the Full Agnostic Regime

The reduction provided by Algorithm 2 admits clean analysis, and nicely harnesses the power of realizable active learners for a single distribution One might wonder if a similar strategy, this time harnessing agnostic learners, might provide a consistent strategy in full agnostic regime. Unfortunately, this is false. It fails even with small amounts of noise localized to the decision boundary, and each h^*_g is the Bayes optimal classifier on D_k. The reason for this lack of consistency comes down to the fact that labeling with the Bayes optimal under-estimates the distorts the noise rates on each target, which in turn biases the output of the ERM step. We present example to this endin the Appendix. We note that any estimate of the noise rates on each group will require G/ϵ^2 label queries, reintroducing dependence on $1/\epsilon^2$.

7.2 A 3ν-Approximation Algorithm

Although the strategy of creating an artificially labeled training set with near-optimal hypotheses on each target fails outside of the target-realizable case, it possesses a nice approximation property.

We give a guarantee to this end in Theorem 3. It states that if we call an active learner with agnostic guarantees on each group D_g, and then use the outputs $\hat{h}^*_g$ to artificially label a new batch of unlabeled data from each group, using ERM on this artificially labeled data gives at worst a $2\nu+\epsilon$ suboptimal hypothesis with high probability. Recall that ν is the noise rate on the multi-group objective. The proof is very similar to that of Theorem 2, but notes in addition that $\hat{h}^*_g$ mislabels on a roughly $\nu_g := \inf_{h\in\mathcal{H}} L\mathcal{G}(h \mid g)$ fraction of the unlabeled samples from each group G when constructing the artificially labeled set.

Theorem 3. . *Fix an arbitrary $\mathcal{H}$ with $d<\infty$, and an agnostic active learner $\mathcal{A}$ with the property that for every D and for all settings of ϵ' and δ', with probability $\geq 1-\delta'$ over its queries, its output h' satisfies*

$$L_D(h') \leq \inf_{h\in\mathcal{H}} L_D(h) + \epsilon',$$

using label queries $\leq l(\epsilon',\delta',\mathcal{H},\mathcal{D})$. Then for all settings of ϵ,δ, and for all collections of targets $\mathcal{G}$, with probability $\geq 1-\delta$, the output $\hat{h}$ of Algorithm 2 run with $\mathcal{A}$ satisfies

$$L^{\max}_{\mathcal{G}}(\hat{h}) \leq \inf_{h\in\mathcal{H}} L^{\max}_{\mathcal{G}}(h) + 2\max_{k\in[D]}\nu_k + \epsilon \leq 3\inf_{h\in\mathcal{H}} L^{\max}_{\mathcal{G}}(h) + \epsilon.$$

and, the label complexity bounded above by

$$O\left(\sum_{G\in\mathcal{G}} l(\epsilon,\delta,\mathcal{H},\mathcal{D}_G)\right).$$

8 Conclusion

In conclusion, we hope that we make people think this is an interesting problem and get some insights into it.

- theory questions: is the power of active learning in this setting mostly held in being able to sample from groups themselves? is the factor of ϵ^2 necessary even in lower noise settings? are there futher itermediate regimes between full agnostic and group realizable where exponential speedup possible?
- is agnostic learning harder or easier than proper learning
- practice questions: can we dsign executable algoithms.

References

[1] Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[2] Huy L. Nguyen and Lydia Zakynthinou. Improved algorithms for collaborative pac learning, 2018.

[3] Nika Haghtalab, Michael I. Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions, 2023.

[4] Jacob Abernethy, Pranjal Awasthi, Matthäus Kleindessner, Jamie Morgenstern, Chris Russell, and Jie Zhang. Active sampling for min-max fairness, 2022.

[5] Agnieszka Słowik and Léon Bottou. On distributionally robust optimization and data rebalancing. In *Proc. AIStats 2022*, Feb 2022. to appear.

[6] Christopher Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. *CoRR*, abs/2112.12181, 2021.

[7] Guy N. Rothblum and Gal Yona. Multi-group agnostic PAC learnability. *CoRR*, abs/2105.09989, 2021.

[8] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *CoRR*, abs/1909.12475, 2019.

[9] Steve Hanneke. A bound on the label complexity of agnostic active learning. In *Proceedings of the 24th International Conference on Machine Learning*, page 353–360, 2007.

[10] Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, 2007.

[11] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers?, 2018.

[12] Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust language modeling, 2019.

[13] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019.

[14] Burr Settles. Active learning literature survey. 2009.

[15] Sanjoy Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19):1767–1781, 2011.

[16] Steve Hanneke and Liu Yang. Minimax analysis of active learning. 2014.

[17] Maria-Florina Balcan and Ruth Urner. Active learning-modern learning theory., 2016.

[18] Maria-Florina Balcan, Alina Beygelzimer, and John Langford. Agnostic active learning. In *Proceedings of the 23rd International Conference on Machine Learning*, page 65–72, 2006.

[19] Rui M Castro and Robert D Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 54(5):2339–2353, 2008.

[20] Stanislav Minsker. Plug-in approach to active learning. *Journal of Machine Learning Research*, 13(1), 2012.

[21] Chicheng Zhang and Kamalika Chaudhuri. Beyond disagreement-based agnostic active learning. *CoRR*, abs/1407.2657, 2014.

[22] Alina Beygelzimer, Sanjoy Dasgupta, and John Langford. Importance weighted active learning. *CoRR*, abs/0812.4952, 2008.

[23] Steve Hanneke and Liu Yang. Negative results for active learning with convex losses. *Journal of Machine Learning Research - Proceedings Track*, 9:321–325, 01 2010.

[24] Matti Kääriäinen. Active learning in the non-realizable case. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2006.

[25] Steve Hanneke. Theory of active learning. 2014.

[26] Chicheng Zhang and Kamalika Chaudhuri. Active learning from weak and strong labelers. *CoRR*, abs/1510.02847, 2015.

[27] David A. Cohn, Les E. Atlas, and Richard E. Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[28] Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[29] Wassily Hoeffding. *Probability Inequalities for sums of Bounded Random Variables*. Springer New York, 1994.

9 Appendix

9.1 Guarantees for General Agnostic Algorithm

We first extend the notation of measure in a slight abuse of notation.

Definition 1. *Given group distribution D_g and a set $S \subseteq \mathcal{X}$ for which $S \cap supp(D_g)$ is measurable under μ_g, define*

$$\mu_g(S) := \mu_g(S \cap supp(D_g)).$$

Definition 2. *Given a collection of group distributions $\mathcal{G}$, and a set $S \subseteq \mathcal{X}$, we say S is "measurable with respect to $\mathcal{G}$" if $S \cap supp(D_g)$ is measurable under μ_g for each $g \in [G]$.*

Definition 3. *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 $L_\emptyset(h) := 1$.

Definition 4. *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 $h(x) = h'(x)$.*

Definition 5. *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 $\mathcal{G}$ and for which $\mathcal{H}'$ agrees on R^c. Given sets of pairs $\mathcal{S}_{R,g}$ and $\mathcal{S}_{R^c,g}$, let*

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

Definition 6. *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 $\mathcal{G}$, and sample sizes $m, m' > 0$, define the function*

$$\Gamma_g(\delta, R, m, m') := \begin{cases} \mu_g(R)\left(\frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(4em/\delta)}{m}}\right) + \sqrt{\frac{\ln(4/\delta)}{2m'}} & \text{if } \mu_g(R) > 0, \mu_g(R^c) > 0 \\ \frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(4em/\delta)}{m}} & \text{if } \mu_g(R) > 0, \mu_g(R^c) = 0 \\ \sqrt{\frac{\ln(4/\delta)}{2m'}} & \text{if } \mu_g(R) = 0, \mu_g(R^c) > 0. \end{cases}$$

Lemma 1. *Fix $\delta \in (0,1)$, and a group distribution $D_g \in \mathcal{G}$ arbitrarily. Further, fix a subset $R \subseteq \mathcal{X}$ measurable with respect to $\mathcal{G}$, 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 $U_g(R \cap supp(D_g))$, and $m' > 0$ samples from $U_g(R^c \cap supp(D_g))$. 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 $R \cap supp(D_g) = \emptyset$ or $R^c \cap supp(D_g) = \emptyset$, 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 $\gamma \in (0,1)$, if $m \geq \frac{32\mu_g(R)^2}{\gamma^2}\left(2d\ln(16/\gamma) + \ln(8/\delta)\right)$ 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 $\mu_g(R \cap supp(D_g)) \neq 0$ and $\mu_g(R^c \cap supp(D_g)) \neq 0$. 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 $U_g(R \cap supp(D_g))$ and their labeling via $O_g(\cdot)$ that we have simultaneously for each $h \in \mathcal{H}'$ that

$$\left|L_{\mathcal{S}_{R,k}}(h) - \mathbb{P}_{(x,y)\sim D_g}\left(h(x) \neq y \middle| x \in R \cap supp(D_g)\right)\right| \leq \frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(4em/\delta)}{m}}.$$

In $R^c \cap supp(D_g)$, 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 Hoeffding's inequality [29], we have with probability $\geq 1-\delta/2$ for all $h \in \mathcal{H}'$ simultaneously

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

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

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

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

$$m \geq \mu_g(R)^2\frac{32}{\gamma^2}\left(2d\ln(16/\gamma) + \ln(8/\delta)\right) > \mu_g(R)^2\frac{32}{\gamma^2}\left(2d\ln(16\mu_g(R)/\gamma) + \ln(8/\delta)\right)$$

implies that

$$\frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(4em/\delta)}{m}} < \frac{\gamma}{2\mu_g(R)}.$$

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

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

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

$$\left|L_{\mathcal{S}_{R,k}}(h) - \mathbb{P}_{(x,y)\sim D_g}\left(h(x) \neq y | \, x \in R \cap supp(D_g)\right)\right| \leq \frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(4em/\delta)}{m}}.$$

When $\mu_g(R^c) = 0$, it holds that $\mu_g(R) = 1$, and so

$$\begin{aligned}
L_{\mathcal{D}}(h \mid g) &= \mathbb{P}_{(x,y)\sim D_g}\left(h(x) \neq y \mid x \in R \cap supp(D_g)\right)\mu_g(R) \\
&\qquad + \mathbb{P}_{(x,y)\sim D_g}\left(h(x) \neq y \mid x \in R^c \cap supp(D_g)\right)\mu_g(R^c) \\
&= \mathbb{P}_{(x,y)\sim D_g}\left(h(x) \neq y \mid x \in R \cap supp(D_g)\right) \\
&\leq L_{\mathcal{S}_{R,g}}(h) + \sqrt{\left(\ln(8/\delta) + d\ln(4em/\delta)\right)/m} \\
&= L_{\mathcal{S};R}(h \mid g) + \Gamma_g(\delta, R, m, m'),
\end{aligned}$$

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

$$m \geq \mu_g(R)^2\frac{32}{\gamma^2}\left(2d\ln(32/\gamma) + \ln(8/\delta)\right) = \frac{32}{\gamma^2}\left(2d\ln(32/\gamma) + \ln(8/\delta)\right)$$

implies that

$$\frac{1}{m} + \sqrt{\frac{\ln(8/\delta) + d\ln(4em/\delta)}{m}} < \frac{\gamma}{2},$$

which by the definition of $\Gamma_g(\delta, R, m, m')$ when $\mu_g(R^c) = 0$ gives us $\Gamma_g(\delta, R, m, m') < \gamma/2 < \gamma$. The case where $\mu_g(R) = 0$ follows the previous argument for when $\mu_g(R^c) = 0$.

□

Definition 7. *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 $\mathcal{G}$, 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*

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

Lemma 2. *Fix $\delta \in (0,1)$, a subset $R \subseteq \mathcal{X}$ measurable with respect to $\mathcal{G}$, 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 $U_g(R \cap supp(D_g))$, and $m'_g > 0$ samples from $U_g(R^c \cap supp(D_g))$. 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 $R \cap supp(D_g) = \emptyset$ or $R^c \cap supp(D_g) = \emptyset$, then we set the corresponding sample to be $\emptyset$. Then with probability $\geq 1 - \delta$, it holds for all $h \in \mathcal{H}'$ that*

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

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

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

Thus we may write

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

□

Definition 8. *Give a set of hypotheses $\mathcal{H}' \subseteq \mathcal{H}$, the "disagreement region of $\mathcal{H}'$" is the set*

$$\Delta(\mathcal{H}') := \{x \in \mathcal{X} : \exists h, h' \in \mathcal{H}' \; s.t. \; h(x) \neq h'(x)\}.$$

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[1] [2]

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

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 $|L^{\max}_{\mathcal{S};R_i}(h) - L^{\max}_{\mathcal{D}}(h)| \leq 2^{I-i}\epsilon/8$, we may write

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

where the first inequality comes from the optimality of h^*. □

[1] 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 sligthtly more samples than Lemma 1 directly suggests.

[2] Because we take the largest measure of the disagreement region over groups as m_i, $m_g \geq$ to the sample size suggested by Lemma 1 for each g.

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*

$$|L_{\mathcal{D}}^{\max}(h) - L_{\mathcal{D}}^{\max}(h^*)| \leq 2^{I-i}\epsilon.$$

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

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

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

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

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

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

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

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

which is the desired result. □

Definition 9. *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 $\rho_g(h, h') := \mu_g\left(h(x) \neq h'(x)\right)$.

Definition 10. *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*

$$\theta_g := \sup_{h \in \mathcal{H}} \sup_{r' \geq r} \frac{\mu_g\left(\mathrm{DIS}(B_g(h, r'))\right)}{r'}.$$

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

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

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

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

and its label complexity is bounded by

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

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 $\left|L_{\mathcal{D}}^{\mathcal{G}}(h) - L_{\mathcal{D}}^{\mathcal{G}}(h^*)\right| \leq 2^{I-i}\epsilon$. Thus, after $I = \lceil \log_2(1/\epsilon) \rceil$ iterations, the output $\hat{h}$ satisfies the consistency condition automatically.

To see the label complexity, we note at iteration i, we label no more than

$$2048\left(\frac{m_i}{\epsilon 2^{I-i}}\right)^2\left(2d\log\left(\frac{128}{\epsilon}\right)+\ln\left(\frac{8G\lceil\log(1/\epsilon)\rceil}{\delta}\right)\right)+\frac{2\ln(4/\delta)}{\epsilon^2}$$

samples for each group distribution D_g, where $m_i = \max_{g'} \mu_{g'}(\Delta(\mathcal{H}_i))$. The only term here that depends on i is $\frac{m_i}{\epsilon 2^{I-i}}$. Note that when $|L^{\max}_{\mathcal{D}}(h) - L^{\max}_{\mathcal{D}}(h^*)| \leq 2^{I-i+1}\epsilon$ - which is true for each $h \in \mathcal{H}_i$ at all iterations i with probability $\geq 1-\delta$ by Lemma 4 - it holds for each $g \in [G]$ that

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

where we recall ν is the noise rate on the multi-group objective. In other words, $h \in B_g(h^*, 2\nu + 2^{I-i+1}\epsilon)$. Thus, with probability $\geq 1-\delta$, for each $g \in [G]$, 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

$$\mu_g(\Delta(\mathcal{H}_i) \cap supp(D_g)) \leq \mu_g(\Delta(B_k(h^*, 2\nu + 2^{I-i+1}\epsilon)) \cap supp(D_g)),$$

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. This allows us to bound the sum of terms depending on i for each distribution D_g as

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

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

9.2 Group-Realizable Guarantees

Theorem 5. *Suppose Algorithm 2 is run with the active learner $\mathcal{A}_{CAL}$ of [27]. 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*

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

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 of [15]. By a union bound , we have that for all $g \in [G]$, $\mathcal{A}_{CAL}$ returns $\hat{h}_g$ with the property that

look up the original source for this bound

$$L_{\mathcal{G}}(\hat{h}_g \mid g) \leq \epsilon/4.$$

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 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

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

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

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

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:

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

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

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

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

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

□

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]. As before, we have that for all $g \in [G]$, $\mathcal{A}_{DHM}$ returns $\hat{h}_g$ with the property that

maybe need an extra factor of 1/2 on delta to make union bound work

$$L_{\mathcal{G}}(\hat{h}_g \mid g) \leq \nu_g + \epsilon/4.$$

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:

make sure you check these guarantees with DHM and make sure you're using them correctly. I think so but im sick so need to come back

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

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

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

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

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

□