Group-realizable multi-group learning by minimizing empirical risk

Navid Ardeshir NAVID.ARDESHIR@COLUMBIA.EDU
Columbia University

Samuel Deng SAMDENG@CS.COLUMBIA.EDU
Columbia University

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Columbia University

Jingwen Liu JINGWENLIU@CS.COLUMBIA.EDU
Columbia University

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

The sample complexity of multi-group learning is shown to improve in the group-realizable setting over the agnostic setting, even when the family of groups is infinite so long as it has finite VC dimension. The improved sample complexity is obtained by empirical risk minimization over the class of group-realizable concepts, which itself could have infinite VC dimension. Implementing this approach is also shown to be computationally intractable, and an alternative approach is suggested based on improper learning.

Keywords: Multi-group learning, group-realizability, empirical risk minimization, sample complexity, computational complexity

1. Introduction

Multi-group learning (Rothblum and Yona, 2021) extends the basic framework of statistical learning to study the performance of predictive models within families of subpopulations. Although multi-group learning can be cast as a special case of other learning frameworks that address subpopulation-level criteria (e.g., Hebert-Johnson et al., 2018; Kim et al., 2019; Dwork et al., 2021; Haghtalab et al., 2023), the goal of this work is to study a natural assumption within multi-group learning—*group-realizability*—under which one might expect improved statistical efficiency and/or computational efficiency as compared to the general case.

In multi-group learning (for binary classification), subpopulations are specified by subsets of the input domain $\mathcal{X}$, and the learning objective concerns a family of (possibly overlapping) subpopulations $\mathcal{G} \subseteq 2^{\mathcal{X}}$. In particular, for a given collection of benchmark classifiers $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ and an excess error rate bound $\epsilon \in (0,1)$, the learner seeks to construct a classifier $f\colon \mathcal{X} \to \{-1,1\}$ using an i.i.d. sample from a probability distribution D over $\mathcal{X} \times \{-1,1\}$ so that, with high probability,

$$\operatorname{err}(f \mid g) \leq \inf_{h \in \mathcal{H}} \operatorname{err}(h \mid g) + \epsilon \quad \text{for each subpopulation } g \in \mathcal{G}. \tag{1}$$

Above, $\operatorname{err}(f \mid g) := \Pr_{(\mathbf{x},\mathbf{y})\sim D}[f(\mathbf{x}) \neq \mathbf{y} \mid \mathbf{x} \in g]$ is the *conditional error rate of a classifier f given g* (defined whenever g has non-zero mass under the marginal of D over $\mathcal{X}$). Importantly, each subpopulation g may have a different optimal benchmark classifier $h^*_g \in \mathcal{H}$, and it is possible that these per-group optimal classifiers have disagreements $h^*_g(x) \neq h^*_{g'}(x)$ at points of intersection

$x \in g \cap g'$. Because of this, it is possible that no $f \in \mathcal{H}$ can satisfy (1), and existing multi-group learning algorithms instead construct f as an ensemble classifier involving functions from $\mathcal{H}$ and $\mathcal{G}$.

Group-realizability is the assumption on $(\mathcal{G}, \mathcal{H}, D)$ that, for each $g \in \mathcal{G}$, there is a benchmark classifier $h^*_g \in \mathcal{G}$ such that $\mathrm{err}(h^*_g \mid g) = 0$. Under this assumption, the goal in multi-group learning is to construct a classifier f such that $\mathrm{err}(f \mid g) \leq \epsilon$ for each $g \in \mathcal{G}$. Group-realizability should be contrasted with the standard *realizability* (or *separability*) assumption on $(\mathcal{H}, D)$, i.e., the existence of $h^* \in \mathcal{H}$ such that $\Pr_{(\mathbf{x},\mathbf{y})\sim D}[h^*(\mathbf{x}) \neq \mathbf{y}] = 0$. Realizability implies group-realizability, but the reverse implication need not hold (unless, e.g., $\mathcal{X} \in \mathcal{G}$).

A marked difference between the "realizable" and "agnostic" (i.e., non-realizable) settings in (single-group) statistical learning comes in the worst-case dependence on ϵ in the sample complexity: roughly $1/\epsilon$ versus $1/\epsilon^2$ (ignoring the dependence on $\mathcal{H}$). A similar sample complexity difference also manifests in multi-group learning with and without the group-realizability assumption, as shown by Tosh and Hsu (2021) for the case where the family of subpopulations $\mathcal{G}$ is finite. Moreover, the computational complexity is, roughly speaking, no worse than that of finding a consistent classifier in $\mathcal{H}$ for each subpopulation $g \in \mathcal{G}$. This is notable since, for some classes such as half-spaces, finding a consistent classifier can be done in polynomial-time, whereas finding a classifier in the class with approximately-minimal error rate on a sample may be computationally intractable (Feldman et al., 2009; Guruswami and Raghavendra, 2009).

In this work, we show that the improved sample complexity (up to a logarithmic factor in $1/\epsilon$) for multi-group learning under group-realizability extends to the case where $\mathcal{G}$ is infinite but has finite VC dimension. To achieve this sample complexity guarantee, we introduce the class $\mathcal{C}_{\mathcal{G},\mathcal{H}} \subseteq \{-1,1\}^{\mathcal{X}}$ of *group-realizable concepts*: the set of functions $c\colon \mathcal{X} \to \{-1,1\}$ consistent with the group-realizability assumption. With this definition in hand, our main result follows simply from *empirical risk minimization (ERM)* over $\mathcal{C}_{\mathcal{G},\mathcal{H}}$. In particular, there is no explicit or algorithmic regularization (in contrast to the algorithm of Tosh and Hsu (2021), which involves a form of aggregation). Remarkably, it is possible for this class $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ to have infinite VC dimension, even when both $\mathcal{G}$ and $\mathcal{H}$ have finite VC dimension. ERM with infinite VC dimension classes is not generally an effective learning procedure in the standard statistical learning setup. However, in our setup, the efficacy of ERM with such a class comes from the restricted sense in which the learned classifier is evaluated. This is very simply captured using shattering coefficients. The statistical efficiency of the $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ class is reminiscent of the near-optimality of ERM in realizable (single-group) statistical learning, where search over $\mathcal{H}$ itself is sufficient for similar sample complexity guarantees.

Unfortunately, the rub with using ERM—which ultimately boils down to just finding a function $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ consistent with the training data—is computational intractability: it generally involves solving an NP-hard problem. The hardness does not come from any potential intractability of optimizing over $\mathcal{H}$ or even enumerating $\mathcal{G}$: our proof of hardness in Section 4 goes via a reduction to instances where $\mathcal{G}$ has polynomial-size and optimization over $\mathcal{H}$ can be performed in polynomial-time. Rather, the difficulty appears to come from even specifying a classifier in $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ when one can only refer to classifiers from $\mathcal{H}$ and subpopulations from $\mathcal{G}$. We show, in Section 5, how "improper learning" can get around this intractability in certain cases.

Relation to prior works

The literature on multi-group learning has focused primarily on the general (agnostic) setting, whether in the online setting (Blum and Lykouris, 2020; Deng et al., 2024) or the batch setting (Roth-

blum and Yona, 2021; Tosh and Hsu, 2021; Rittler and Chaudhuri, 2023; Deng and Hsu, 2024). The work of Tosh and Hsu (2021) provides a (batch) multi-group learning algorithm, called "Prepend", for families of subpopulations $\mathcal{G}$ that may be infinite but have finite VC dimension. Prepend—which is nearly identical to an algorithm proposed by Globus-Harris et al. (2022) in a somewhat related context—is efficient as long as it has access to an oracle for solving ERM-type optimization problems over $\mathcal{H} \times \mathcal{G}$. Such "oracle efficient" algorithms have a long history in many areas of learning theory (e.g., Kalai and Vempala, 2005; Kakade and Kalai, 2005; Dasgupta et al., 2007; Dudík et al., 2011; Dann et al., 2018; Foster and Rakhlin, 2020; Haghtalab et al., 2022; Wang et al., 2022; Garg et al., 2024), including the closely related subject of subgroup fairness (Kearns et al., 2018).

An important drawback of Prepend, however, is that its sample complexity is suboptimal: thr dependence on ϵ is roughly $1/\epsilon^3$. (Here, for simplicity, we are omitting the dependence on $\mathcal{G}$, $\mathcal{H}$, and the smallest group probability mass.) Prepend can be specialized to the group-realizable setting, but its sample complexity remains suboptimal: roughly $1/\epsilon^2$. (Another algorithm of Kim et al. (2019) based on multi-accuracy also has a suboptimal $\Omega(1/\epsilon^2)$ sample complexity under group-realizability.) As mentioned before, Tosh and Hsu (2021) do provide an algorithm (different from Prepend) with near-optimal dependence on ϵ in the sample complexity (in both the agnostic case and under group-realizability) for the case where $\mathcal{G}$ is finite. The sample complexity of their algorithm is linear in $\log(|\mathcal{G}|)$; the algorithm also explicitly enumerates $\mathcal{G}$, and hence the computational complexity may be exponential in the sample size. No other prior works on multi-group learning address the potential sample complexity improvements in the group-realizable setting, including works that view multi-group learning as a special case of other learning frameworks such as multi-calibration (Hebert-Johnson et al., 2018), outcome indistinguishability (Dwork et al., 2021; Rothblum and Yona, 2021), and multi-objective learning (Haghtalab et al., 2023). Consequently, sample complexity guarantees obtained by reducing to these general frameworks are $1/\epsilon^2$ or worse.

Interestingly, our identification of $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ as the correct object of study leads to our main sample complexity result via a simple analysis. The introduction of this class of functions $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ does not appear in prior literature. Instead, prior algorithms in multi-group learning such as those presented in Tosh and Hsu (2021) and Rothblum and Yona (2021) as noted above consider aggregation procedures over $\mathcal{H}$ and $\mathcal{G}$ instead of explicitly defining the class of interest to search over.

The NP-hardness of the computational problem encountered by ERM over $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ is reminiscent of the hardness of proper learning in PAC learning (e.g., Pitt and Valiant, 1988), although as discussed above, the nature of the hardness appears to be conceptually different.

2. Setting

Throughout, $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}} \cong 2^{\mathcal{X}}$ refers to a family of subpopulations (a.k.a. groups), and $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ refers to a collection of benchmark classifiers (a.k.a. hypotheses). It will be convenient to regard a group $g \in \mathcal{G}$ both as a function $g\colon \mathcal{X} \to \{0,1\}$ and as a subset $g \subseteq \mathcal{X}$. For a class of functions $\mathcal{F}$ defined over a domain $\mathcal{Z}$, the k-th shattering coefficient is the largest possible number of k-tuples realized by $\mathcal{F}$ on k points from $\mathcal{Z}$:

$$\mathcal{S}_{\mathcal{Z}}(\mathcal{F}, k) := \sup_{z_1,\ldots,z_k \in \mathcal{Z}} |\{(f(z_1),\ldots,f(z_k)) \mid f \in \mathcal{F}\}|.$$

A central contribution of this paper is the introduction of the class of *group-realizable concepts* $\mathcal{C}_{\mathcal{G},\mathcal{H}}$, which we define as follows.

Definition 1 *Fix any family of groups $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ and any hypothesis class $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ on the same input domain $\mathcal{X}$. The set of* group-realizable concepts with respect to $(\mathcal{G}, \mathcal{H})$*, denoted by $\mathcal{C}_{\mathcal{G},\mathcal{H}} \subseteq \{-1,1\}^{\mathcal{X}}$, is the set of all functions $c\colon \mathcal{X} \to \{-1,1\}$ such that, for each $g \in \mathcal{G}$, there exists $h \in \mathcal{H}$ satisfying $c(x) = h(x)$ for all $x \in g$. Additionally, for a probability distribution D on $\mathcal{X} \times \{-1,1\}$, we say $(\mathcal{G}, \mathcal{H}, D)$ satisfies* group-realizability *if there exists $c^* \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ such that $\Pr_{(\mathbf{x},\mathbf{y})\sim D}[c^*(\mathbf{x}) = \mathbf{y}] = 1$.*

In this work, we only consider distributions D over $\mathcal{X} \times \{-1,1\}$ such that $(\mathcal{G}, \mathcal{H}, D)$ satisfies group-realizability. So we can equivalently specify D by its marginal distribution P over $\mathcal{X}$, and a group-realizable concept $c^* \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$. The conditional error rate of a classifier $f\colon \mathcal{X} \to \{-1,1\}$ given g can therefore be written as

$$\operatorname{err}(f \mid g) = \Pr_{\mathbf{x}\sim P}[f(\mathbf{x}) \neq c^*(\mathbf{x}) \mid \mathbf{x} \in g].$$

Our concern is the *distribution-free* setting in the sense that we are interested in guarantees that hold for worst-case choices of P. This mirrors the usual sense in which standard PAC learning is regarded as distribution-free even under realizability (Valiant, 1984). Our aim is to achieve complexity guarantees that are comparable to those achievable in the standard setting when $\mathcal{G} = \{\mathcal{X}\}$. Formally, our goal is to furnish a classifier $f\colon \mathcal{X} \to \{-1,1\}$ whose conditional-error rate on each group is small, captured by the following definition.

Definition 2 *For any $(\mathcal{G}, \mathcal{H}, D)$ satisfying group-realizability (Definition 1), a classifier $f\colon \mathcal{X} \to \{-1,1\}$ achieves* group-realizable multi-group learning *if, for all $(\epsilon, \delta) \in (0,1)$,*

$$\operatorname{err}(f \mid g) \leq \epsilon \quad \text{for all } g \in \mathcal{G}$$

with probability $1-\delta$ over the i.i.d. training examples $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \sim P^n$.

Specifying a group-realizable concept $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ typically begins by specifying an assignment $\mathcal{G} \ni g \mapsto h_g \in \mathcal{H}$ of hypotheses to groups. Note that an assignment of hypotheses to groups for which there are "disagreements" (e.g., $h_g(x) \neq h_{g'}(x)$ for some $x \in g \cap g'$) does not directly yield a well-defined classifier. In this work, we consider the following options to get a valid classifier:

1. Ensure the chosen h_g's have no disagreements, i.e., whenever any two groups $g, g' \in \mathcal{G}$ have a non-empty intersection, we have $h_g(x) = h_{g'}(x)$ for all $x \in g \cap g'$, so the assignment corresponds to some $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$. The sample and computational complexities of this approach are explored in Sections 3 and 4.

2. Reconcile disagreements among the h_g's; a data-driven approach is given in Section 5. This approach is not guaranteed to yield a concept from $\mathcal{C}_{\mathcal{G},\mathcal{H}}$.

Even when $\mathcal{G}$ and $\mathcal{H}$ have finite VC dimension, it is possible for $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ to have infinite VC dimension, as the following proposition shows.

Proposition 3 *There exists $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ and $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ on the same input domain $\mathcal{X}$, each with finite VC dimension, such that $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ has infinite VC dimension.*

Proof Let $\mathcal{X}$ be any set of infinite cardinality (e.g., the integers). Let $\mathcal{H}$ consist of just the "constant 1" function and the "constant -1" function. This class is finite (and has VC dimension 1). Let $\mathcal{G} = \{\{x\} \mid x \in \mathcal{X}\}$ be the family of singleton sets. This class has VC dimension 1. The groups in $\mathcal{G}$ are disjoint, so $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ contains all $\{-1,1\}$-valued functions on $\mathcal{X}$. Therefore, finite subsets of $\mathcal{X}$ of all sizes are shattered by $\mathcal{C}_{\mathcal{G},\mathcal{H}}$, which means that $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ has infinite VC dimension. ■

The set of group-realizable concepts $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ is not an appropriate class to use in the agnostic setting, where $(\mathcal{G},\mathcal{H},D)$ may fail to satisfy group-realizability. For instance, suppose $\mathcal{G} = \{g_1, g_2\}$ with $g_1 \cap g_2 \neq \emptyset$, and $\mathcal{H} = \{x \mapsto -1, x \mapsto 1\}$ contains just the constant -1 and constant 1 hypotheses. Then $\mathcal{C}_{\mathcal{G},\mathcal{H}} = \{x \mapsto -1, x \mapsto 1\}$ as well. However, consider $(\mathbf{x},\mathbf{y}) \sim D$ with $P(g_1 \setminus g_2) = P(g_1 \cap g_2) = P(g_2 \setminus g_1) = 1/3$ (where P is the marginal distribution of $\mathbf{x}$), $\Pr[\mathbf{y} = 1 \mid \mathbf{x} \in g_1 \setminus g_2] = 1/2$, $\Pr[\mathbf{y} = 1 \mid \mathbf{x} \in g_1 \cap g_2] = 2/3$, $\Pr[\mathbf{y} = 1 \mid \mathbf{x} \in g_2 \setminus g_1] = 0$. In this case, the best constant predictor for g_1 is $x \mapsto 1$, but the best constant predictor for g_2 is $x \mapsto -1$.

3. Sample complexity

Our main sample complexity result is a consequence of the following theorem.

Theorem 4 *Let P be a probability distribution on $\mathcal{X}$, let $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$ be any family of groups on $\mathcal{X}$, and let $\mathcal{H} \subseteq \{-1,1\}^{\mathcal{X}}$ be any hypothesis class on $\mathcal{X}$. Fix any $c^* \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$, and let $(\mathbf{x}_1,\ldots,\mathbf{x}_n) \sim P^n$ and $\mathbf{S} = ((\mathbf{x}_i, c^*(\mathbf{x}_i)))_{i\in[n]}$.*

1. *For any $g \in \mathcal{G}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, every $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ consistent with $\mathbf{S}$ has*

$$\Pr_{\mathbf{x}\sim P}[c(\mathbf{x}) \neq c^*(\mathbf{x}) \wedge \mathbf{x} \in g] \leq \frac{4\left(\log\left(\binom{2n}{\leq d_{g,\mathcal{H}}}\right) + \log\left(\frac{4}{\delta}\right)\right)}{n}. \tag{2}$$

   Above, $d_{g,\mathcal{H}}$ is the VC dimension of $\mathcal{H}$ restricted to $g \subseteq \mathcal{X}$.

2. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, every $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ consistent with $\mathbf{S}$ has*

$$\Pr_{\mathbf{x}\sim P}[c(\mathbf{x}) \neq c^*(\mathbf{x}) \wedge \mathbf{x} \in g] \leq \frac{4\left(\log\left(\binom{2n}{\leq d_{\mathcal{G}}}\right) + \log\left(\binom{2n}{\leq \sup_{g\in\mathcal{G}} d_{g,\mathcal{H}}}\right) + \log\left(\frac{4}{\delta}\right)\right)}{n} \quad \forall g \in \mathcal{G}. \tag{3}$$

   Above, $d_{g,\mathcal{H}}$ is the VC dimension of $\mathcal{H}$ restricted to $g \subseteq \mathcal{X}$, and $d_{\mathcal{G}}$ is the VC dimension of $\mathcal{G}$.

Although the guarantees of Theorem 4 are not stated in terms of the conditional error rates $\mathrm{err}(c \mid g)$, such error rates can be gotten by dividing by $P(g) = \Pr_{\mathbf{x}\sim P}[\mathbf{x} \in g]$. After doing so, the right-hand sides in (2) and (3) have a denominator of $nP(g)$, which can be interpreted as the expectation of number of training examples N_g from g. Therefore, the conditional error rate decreases roughly as $1/N_g$, which should be contrasted to the $1/\sqrt{N_g}$ rates in the agnostic case (Tosh and Hsu, 2021).

Furthermore, to obtain sample size requirements for multi-group learning, we can simply "solve for n" to make the right-hand side equal to (or bounded above by) the target excess error rate bound ϵ. For instance, from (3), we derive the sample size requirement

$$n \geq C \cdot \frac{(d_{\mathcal{G},\mathcal{H}} + d_{\mathcal{G}})\log(1/\gamma\epsilon) + \log(1/\delta)}{\gamma\epsilon}, \tag{4}$$

where $C > 0$ is some absolute constant, $d_{\mathcal{G},\mathcal{H}} := \sup_{g\in\mathcal{G}} d_{g,\mathcal{H}}$, and γ is a lower-bound on $\Pr_{\mathbf{x}\sim P}[\mathbf{x} \in g]$ that holds for all $g \in \mathcal{G}$.[1] Note that $d_{\mathcal{G},\mathcal{H}}$ is always at most the VC dimension $d_{\mathcal{H}}$ of $\mathcal{H}$, and is equal to $d_{\mathcal{H}}$ whenever $\mathcal{X} \in \mathcal{G}$. For comparison, the sample size requirement of the algorithm of Tosh and Hsu (2021) in the group-realizable setting is

$$n \geq C \cdot \frac{d_{\mathcal{H}} \log(1/\gamma\epsilon) + \log(|\mathcal{G}|) + \log(1/\delta)}{\gamma\epsilon}. \tag{5}$$

(See Section 5 for more discussion.) In both cases, when $\mathcal{G} = \{\mathcal{X}\}$, the sample size requirement reduces to the usual sample complexity for ERM in the realizable setting (which is within a factor of $\log(1/\epsilon)$ from optimal). Because multi-group learning generalizes classical (single-group) realizable learning when $\mathcal{G} = \{\mathcal{X}\}$, the classical lower bound of $\Omega(d_{\mathcal{H}}/\epsilon)$ of Blumer et al. (1989) demonstrates that our dependence on ϵ is tight up to the $\log(1/\epsilon)$ factor.

The difference between (4) and (5) manifests for classes where $\log(|\mathcal{G}|) \gg d_{\mathcal{G}} \log(1/\gamma\epsilon)$. Perhaps more interesting, however, is that (4) is achieved via ERM over $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ without any explicit regularization, which stands in contrast to the algorithm of Tosh and Hsu (2021), which explicitly uses aggregation.

In the first part of Theorem 4 (specifically (2)), we see that there is no dependence on $\mathcal{G}$ whatsoever. The guarantee in (2) holds with probability $1 - \delta$ for each group $g \in \mathcal{G}$, but not for all groups simultaneously. This is relevant in cases where the downstream evaluation is ultimately based only on performance in a single group, but the identity of that group is not known at the training time.

The proof of (each part of) Theorem 4 is a simple consequence of the following lemma. In Lemma 5, we use the standard shorthand $Pf := \mathbb{E}_{\mathbf{x}\sim P}[f(\mathbf{x})]$ and $P_n f := \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)$ for random variables and $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \sim P^n$.

Lemma 5 (Vapnik and Chervonenkis, 1971) *Let $\mathcal{F}$ be a family of measurable functions $f\colon \mathcal{Z} \to \{0,1\}$, and let $\delta \in (0,1)$. Let $\alpha_n = (4/n)\ln(4\mathcal{S}_{\mathcal{Z}}(\mathcal{F}, 2n)/\delta)$, where $\mathcal{S}_{\mathcal{Z}}(\mathcal{F}, k)$ is the k-th shattering coefficient for the class $\mathcal{F}$. Let P be any distribution over $\mathcal{Z}$, and let P_n be the empirical distribution on an i.i.d. sample of size n from P. With probability at least $1 - \delta$, for all $f \in \mathcal{F}$:*

$$\frac{Pf - P_n f}{\sqrt{Pf}} \leq \sqrt{\alpha_n}$$

Proof of Theorem 4 For the first part of the claim, we let

$$\mathcal{F} := \{x \mapsto g(x)(c\triangle c^*)(x) \mid c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}\},$$

where $(c\triangle c^*)(x) = \mathbb{1}\{c(x) \neq c^*(x)\}$. Consider any $2n$ points $X := (x_1, \ldots, x_{2n})$ from $\mathcal{X}$, so

$$\mathcal{F}|_X = \left\{(f(x_i))_{i\in[2n]} \;\middle|\; f \in \mathcal{F}\right\} = \left\{(g(x_i)(c\triangle c^*)(x_i))_{i\in[2n]} \;\middle|\; c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}\right\}.$$

By definition of $\mathcal{C}_{\mathcal{G},\mathcal{H}}$, each vector $v \in \mathcal{F}|_X$ is equivalent to $(g(x_i)(h\triangle c^*)(x_i))_{i\in[2n]}$ for some $h \in \mathcal{H}$. Therefore,

$$|\mathcal{F}|_X| \leq \mathcal{S}_g(\mathcal{H}, 2n) \leq \binom{2n}{\leq d_{g,\mathcal{H}}}$$

1. The dependence on γ can be alleviated if one is willing to consider non-uniform error bounds that scale (inversely) with $P(g)$ for group g; see discussion of Tosh and Hsu (2021) on this matter. This aspect is common to all multi-group learning algorithms.

by Sauer's lemma. Since this holds for all choices of $x_1, \ldots, x_{2n} \in \mathcal{X}$, the above bound also holds for $\mathcal{S}_{\mathcal{X}}(\mathcal{F}, 2n)$. To finish the proof, we use a standard manipulation on Lemma 5. For any non-negative numbers A, B, C, we have that $A \leq B + C\sqrt{A}$ implies $A \leq B + C^2 + \sqrt{B}C$, which results in

$$Pf \leq P_n f + 2\sqrt{P_n f \frac{\log \mathcal{S}_{\mathcal{X}}(\mathcal{F}, 2n) + \log(4/\delta)}{n}} + 4\frac{\log \mathcal{S}_{\mathcal{X}}(\mathcal{F}, 2n) + \log(4/\delta)}{n}$$

for all $f \in \mathcal{F}$. We assume that $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ is consistent with $\mathbf{S}$, so noting that $f(x) = g(x)(c \triangle c^*)(x)$ implies that $P_n f = 0$. Therefore,

$$Pf \leq 4\frac{\log \mathcal{S}_{\mathcal{X}}(\mathcal{F}, 2n) + \log(4/\delta)}{n}$$

and the result follows from plugging in the bound on $\mathcal{S}_g(\mathcal{H}, 2n)$ established above.

For the second claim, we just change the class $\mathcal{F}$ to

$$\mathcal{F} := \{x \mapsto g(x)(c \triangle c^*)(x) \mid g \in \mathcal{G}, c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}\}.$$

Again, for any $2n$ points $X := (x_1, \ldots, x_{2n})$ from $\mathcal{X}$,

$$\mathcal{F}|_X = \left\{(g(x_i)(c \triangle c^*)(x_i))_{i \in [2n]} \;\middle|\; g \in \mathcal{G}, c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}\right\}.$$

By definition of $\mathcal{C}_{\mathcal{G},\mathcal{H}}$, each vector $v \in \mathcal{F}|_X$ is equivalent to $(g(x_i)(h \triangle c^*)(x_i))_{i \in [2n]}$ for some $g \in \mathcal{G}$ and some $h \in \mathcal{H}$. Each such vector is obtained (via component-wise product) from a vector of the form $(g(x_i))_{i \in [2n]}$ for some $g \in \mathcal{G}$, and a vector of the form $(h(x_i))_{i \in [2n], x_i \in g}$ for some $h \in \mathcal{H}$. Therefore,

$$|\mathcal{F}|_X| \leq \mathcal{S}_{\mathcal{X}}(\mathcal{G}, 2n) \cdot \sup_{g \in \mathcal{G}} \mathcal{S}_g(\mathcal{H}, 2n) \leq \binom{2n}{\leq d_{\mathcal{G}}} \cdot \sup_{g \in \mathcal{G}} \binom{2n}{\leq d_{g,\mathcal{H}}}.$$

The rest is the same as the proof for the first part. ■

In the proof of Theorem 4, we see that the full "richness" of $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ is never encountered because the "mistake behaviors" that are counted are always restricted to individual groups. On any group $g \in \mathcal{G}$, the behavior of a group-realizable concept $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ is determined by some hypothesis $h_g \in \mathcal{H}$, so the number of behaviors is limited if $\mathcal{G}$ and $\mathcal{H}$ have finite VC dimension. This kind of argument is reminiscent of the concept of computational indistinguishability from cryptography, which has recently been adopted in the context of learning criteria such as multi-calibration (Hebert-Johnson et al., 2018) and outcome indistinguishability (Dwork et al., 2021). In our case, the "adversary" that tries to find a fault in a learner's classifier makes an appearance only in the analysis, rather than explicitly in an algorithm that, say, simulates game dynamics (e.g., Haghtalab et al., 2023).

4. Computational complexity

Even though ERM with $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ has good sample complexity in the group-realizable setting, it may be difficult to implement, even in cases where ERM is easy to implement with $\mathcal{H}$ itself (and where the size of the family of groups is small). Specifically, we show that given a succinct description of $\mathcal{G}$ and $\mathcal{H}$, along with a labeled dataset S, it is NP-hard to decide if there is a concept in $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ that

is consistent with S, even if provided an oracle that returns a consistent hypothesis from $\mathcal{H}$ for any given labeled dataset (should one exist).

We give a reduction from the NP-complete decision problem ONE-IN-THREE 3SAT (Schaefer, 1978; Garey and Johnson, 1979): Given a 3-CNF formula ϕ, decide if there is a truth assignment to the variables such that each clause in ϕ has exactly one literal that evaluates to "true".

Overview of the reduction. Given a 3-CNF formula, we construct a group family with one group per clause, and a hypothesis class where hypotheses correspond to truth assignments. The hypothesis class is a disjoint union of clause-specific (i.e., group-specific) hypothesis classes, where the hypotheses for a particular clause are those that correspond to all possible truth assignments that make *exactly* one literal in the clause evaluate to "true". Recall that to specify a concept in $\mathcal{C}_{\mathcal{G},\mathcal{H}}$, we should choose a hypothesis in $\mathcal{H}$ for each $g \in \mathcal{G}$. Ensuring that the choice of hypothesis for $g \in \mathcal{G}$ comes from the g-specific hypothesis class is handled by introducing additional points in the input domain/groups, and extending the behavior of the hypotheses and constructing the labeled dataset S in a natural way.

Notation. For a literal l, let $v(l)$ denote the variable x such that $l \in \{x, \neg x\}$, and let

$$y(l) := \begin{cases} +1 & \text{if } l = v(l) \\ -1 & \text{if } l = \neg v(l) \end{cases}$$

denote the "polarity label" of that literal.

Reduction. Let ϕ be a 3-CNF formula over variables $x_1, \ldots, x_n$ with clauses $C_1, \ldots, C_m$. Each clause C_i is the disjunction of three literals, $C_i = l_i^1 \vee l_i^2 \vee l_i^3$.

We use ϕ to define the group family $\mathcal{G}$, hypothesis class $\mathcal{H}$, and labeled dataset S as follows.

1. The input domain $\mathcal{X}$ consists of $n + m$ points, one per variable and clause. We use the same names $(x_1, \ldots, x_n, C_1, \ldots, C_m)$ for these points as the variables and clauses.

2. The group family is $\mathcal{G} := \{g_1, \ldots, g_m\}$, with one group per clause.

3. For each clause $C_i = l_i^1 \vee l_i^2 \vee l_i^3$, define $g_i := \{v(l_i^1), v(l_i^2), v(l_i^3), C_i\}$. Also, define $\mathcal{H}_i := \mathcal{H}_i^1 \cup \mathcal{H}_i^2 \cup \mathcal{H}_i^3$, where for each $t \in \{1, 2, 3\}$, $\mathcal{H}_i^t$ contains all possible hypotheses $h \colon \mathcal{X} \to \{-1, 1\}$ satisfying the following:

   (a) $h(v(l_i^t)) = y(l_i^t)$;
   (b) $h(v(l_i^s)) = -y(l_i^s)$ for $s \neq t$;
   (c) $h(C_i) = +1$;
   (d) $h(C_j) = -1$ for all $j \neq i$.

4. The overall hypothesis class is $\mathcal{H} := \mathcal{H}_1 \cup \cdots \cup \mathcal{H}_m$.

5. The labeled dataset is $S := ((C_1, 1), \ldots, (C_m, 1))$.

Soundness and completeness of the reduction are immediate, and a description of $(\mathcal{G}, \mathcal{H}, S)$ can be produced from ϕ in $\mathrm{poly}(n)$ time.

Efficiency of searching for a consistent hypothesis. We claim that it is easy to search for a consistent hypothesis in the hypothesis class $\mathcal{H}$ constructed by the above reduction. (This holds even for the hypothesis class derived from a CNF formula with no clause width restriction.)

Recall that $\mathcal{H}$ is the (disjoint) union of $\mathcal{H}_1, \ldots, \mathcal{H}_m$. We first show how to search for a hypothesis in $\mathcal{H}_j$ (for a fixed $j \in \{1, \ldots, m\}$) that is consistent with any set of labeled example $S' \subseteq (\mathcal{X} \setminus \{C_1, \ldots, C_m\}) \times \{-1, 1\}$. Define the literals $\ell_{(x,-1)} = \neg x$ and $\ell_{(x,+1)} = x$, and define the term T to be the conjunction of literals $\ell_{(x,y)}$ for $(x, y) \in S'$. Then, there is a hypothesis $h \in \mathcal{H}_j$ consistent with S' if and only if there is a truth assignment such that: T is satisfied, and C_j has exactly one literal that evaluates to "true". To search for such a hypothesis: construct a partial truth assignment that satisfies T; if such a partial assignment exists, then check if it can be extended to satisfy exactly one of the literals in C_j. This can be performed with a linear scan over T and C_j.

To see that it is easy to search for a hypothesis in $\mathcal{H}$ that is consistent with a collection of labeled examples $S \subseteq \mathcal{X} \times \{-1, 1\}$, it suffices to explain how to determine which $\mathcal{H}_j$ to search: this ultimately hinges upon which examples of the form (C_i, y) are in S. If there are no examples of the form (C_i, y) in S, then the search is unrestricted. If $(C_i, +1) \in S$, then the search excludes all $\mathcal{H}_j$ for $j \neq i$. If $(C_i, -1) \in S$, then the search excludes $\mathcal{H}_i$. Of course, it is possible that all hypotheses are ultimately excluded, in which case there is clearly no consistent hypothesis.

Implications. The reduction above shows that it is NP-hard to find a $c \in \mathcal{C}_{\mathcal{G},\mathcal{H}}$ consistent with a collection of labeled examples, even if provided an oracle for finding hypotheses in $\mathcal{H}$ consistent with any given $\mathrm{poly}(N)$-many labeled examples from $\mathcal{X} \times \{-1, 1\}$, and even if $|\mathcal{G}| = \mathrm{poly}(N)$, where N represents the dimension or description length of inputs, hypotheses, and groups.

5. Improper multi-group learning under group-realizability

This computational intractability from Section 4 is specific to "proper" multi-group learning under group-realizability; it only applies to situations where one seeks to find a classifier from $\mathcal{C}_{\mathcal{G},\mathcal{H}}$. As previously mentioned, the intractability can be subverted with improper learning, at least in some situations (including the specific scenario from the reduction in Section 4) by using the approach of Tosh and Hsu (2021):

- Let $\mathbf{S}$ be n i.i.d. labeled examples from the distribution $D = (P, c^*)$.
- For each $g \in \mathcal{G}$, let $\hat{h}_g \in \mathcal{H}$ be any hypothesis consistent with the first $n/2$ examples in $\mathbf{S}$.
- Run Tosh and Hsu's Algorithm 2 with group family $\mathcal{G}$, hypotheses $\{\hat{h}_g \mid g \in \mathcal{G}\}$, and learning rate $\eta = 1/2$ on the last $n/2$ examples in $\mathbf{S}$, to obtain the final classifier f. This algorithm is a specific instantiation of an online learning algortihm of Blum and Mansour (2007) combined with online-to-batch conversion.

The classifier f output in the last step is randomized, although a deterministic classifier can be easily obtained with a simple modification to their algorithm (specifically, replacing the algorithm of Blum and Mansour (2007) with a suitable variant of Littlestone and Warmuth (1994)'s Weighted Majority). In either case, the salient point here is that f is not selected from $\mathcal{C}_{\mathcal{G},\mathcal{H}}$; there is no attempt to ensure that hypotheses assigned to groups (from the second step above) "agree" on regions of intersection, so the intractability results from Section 4 are not applicable. Instead, these hypotheses are combined in an ensemble classifier using online learning and online-to-batch conversion.

As mentioned in Section 3, the sample size requirement of this algorithm is given in (5), which is comparable to that of ERM over $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ when $\log(|\mathcal{G}|) \lesssim d_{\mathcal{G}} \log(1/\epsilon)$. This algorithm runs in polynomial-time whenever $\mathcal{G}$ has polynomial cardinality and finding consistent hypotheses from $\mathcal{H}$ can be done in polynomial-time. These aforementioned conditions hold for the $(\mathcal{G}, \mathcal{H})$ constructed in the reduction from Section 4. It is in these scenarios that computational intractability is subverted by improper learning.

6. Discussion and future directions

The statistical efficiency of ERM over the rich class $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ is a remarkable phenomenon, and it seems worthy of further investigation in other settings, including general (agnostic) multi-group learning with a suitable relaxation of $\mathcal{C}_{\mathcal{G},\mathcal{H}}$. In our setting, ERM over $\mathcal{C}_{\mathcal{G},\mathcal{H}}$ is computationally intractable, but an ensemble method sometimes gets around the intractability without a statistical cost. This is reminiscent of convex relaxation and other ways improper learning offer computational speed-ups. It would be interesting to understand if these other approaches are also applicable in our setting.

A problem left open is to find a general oracle-efficient multi-group learning algorithm that achieves the sample complexity from (4) in the group-realizable setting. One possible line of attack is to leverage recent progress on oracle-efficient algorithms in related settings (Deng et al., 2024; Okoroafor et al., 2025).

Acknowledgments

We acknowledge support from the ONR under grant N00014-24-1-2700. SD also acknowledges the support of the Avanessians Doctoral Fellowship for Engineering Thought Leaders and Innovators in Data Science. This work grew out of discussions during the "Modern Paradigms in Generalization" program at the Simons Institute for the Theory of Computing, Berkeley in 2024.

References

Avrim Blum and Thodoris Lykouris. Advancing subgroup fairness via sleeping experts. In *Innovations in Theoretical Computer Science*, 2020.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC RL with rich observations. In *Advances in Neural Information Processing Systems 31*, 2018.

Sanjoy Dasgupta, Daniel Hsu, and Claire Monteleoni. A general agnostic active learning algorithm. In *Advances in Neural Information Processing Systems 20*, 2007.

Samuel Deng and Daniel Hsu. Multi-group learning for hierarchical groups. In *International Conference on Machine Learning*, 2024.

Samuel Deng, Daniel Hsu, and Jingwen Liu. Group-wise oracle-efficient algorithms for online multi-group learning. In *Advances in Neural Information Processing Systems 37*, 2024.

Miroslav Dudík, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Conference on Uncertainty in Artificial Intelligence*, 2011.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Symposium on Theory of Computing*, 2021.

Vitaly Feldman, Parikshit Gopalan, Subhash Khot, and Ashok Kumar Ponnuswami. On agnostic learning of parities, monomials, and halfspaces. *SIAM Journal on Computing*, 39(2):606–645, 2009.

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, 2020.

Michael R Garey and David S Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

Sumegha Garg, Christopher Jung, Omer Reingold, and Aaron Roth. Oracle efficient online multicalibration and omniprediction. In *Symposium on Discrete Algorithms*, 2024.

Ira Globus-Harris, Michael Kearns, and Aaron Roth. An algorithmic framework for bias bounties. In *Conference on Fairness, Accountability, and Transparency*, 2022.

Venkatesan Guruswami and Prasad Raghavendra. Hardness of learning halfspaces with noise. *SIAM Journal on Computing*, 39(2):742–765, 2009.

Nika Haghtalab, Yanjun Han, Abhishek Shetty, and Kunhe Yang. Oracle-efficient online learning for smoothed adversaries. In *Advances in Neural Information Processing Systems 35*, 2022.

Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. In *Advances in Neural Information Processing Systems 36*, 2023.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, 2018.

Sham Kakade and Adam T Kalai. From batch to transductive online learning. In *Advances in Neural Information Processing Systems 18*, 2005.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, 2018.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Conference on AI, Ethics, and Society*, 2019.

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

Princewill Okoroafor, Robert Kleinberg, and Michael P Kim. Near-optimal algorithms for omniprediction. *arXiv preprint arXiv:2501.17205*, 2025.

Leonard Pitt and Leslie G Valiant. Computational limitations on learning from examples. *Journal of the ACM*, 35(4):965–984, 1988.

Nicholas Rittler and Kamalika Chaudhuri. Agnostic multi-group active learning. In *Advances in Neural Information Processing Systems 36*, 2023.

Guy Rothblum and Gal Yona. Multi-group agnostic PAC learnability. In *International Conference on Machine Learning*, 2021.

Thomas J Schaefer. The complexity of satisfiability problems. In *Symposium on Theory of Computing*, 1978.

Christopher Tosh and Daniel Hsu. Simple and near-optimal algorithms for hidden stratification and multi-group learning. *arXiv preprint arXiv:2112.12181v2*, 2021.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

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

Guanghui Wang, Zihao Hu, Vidya Muthukumar, and Jacob D Abernethy. Adaptive oracle-efficient online learning. In *Advances in Neural Information Processing Systems 35*, 2022.