Polarization Through the Lens of Learning Theory

NIKA HAGHTALAB, Microsoft Research New England
MATTHEW O. JACKSON, Stanford University
ARIEL D. PROCACCIA, Carnegie Mellon University

We present a fresh perspective on belief polarization, based on two learning-theoretic models. In the first model, two agents learn from training sets drawn from different distributions that slightly disagree on some labels. In the second model, two agents learn from training sets sampled from the very same distribution, but pay a cost for the complexity of the hypotheses they learn. We show that in both cases, even when the agents are exposed to almost identical sources of information, they can learn hypotheses that disagree substantially. However, in the latter model, we demonstrate that this phenomenon can be alleviated by introducing a slight bias into the information selection process.

1 INTRODUCTION

In 1998, the Lancet, a respected medical journal, published a study linking the MMR vaccine to autism. Although this study has since been thoroughly debunked and retracted, it is still a rallying cry for the modern anti-vaccination movement. Nowadays periodic outbreaks of measles are often associated with pockets of resistance to vaccination — like the recent one in Portland, Oregon, in January 2019 [26]. Research shows that the opinions of Americans about vaccines are highly polarized, but, interestingly, they are not divided along the usual political fault lines; rather, the more political one is (in either direction), the more likely one is to think vaccines are unsafe [14].

The vaccine controversy is just one example of belief polarization, which, of course, is not a new phenomenon. The prevalence of echo chambers is inevitable, since people mostly interact with others who share their background as well as opinions (see [15]), and can even interpret information to confirm what they already believe (e.g., [12]).[1] But this phenomenon has been exacerbated by the advent of cable television, the Internet and social network platforms, which immerse users in content that is tailored to their existing preferences and shared by like-minded people (e.g., see [8]). In this reality, polarization is a natural outcome: when different people are exposed to very different sources of information, they are bound to arrive at different conclusions.

However, we are interested in a more subtle — more insidious — type of polarization: that which occurs when individuals are exposed to *similar* information, yet still end up having substantially different opinions. Our research question is this: Under what conditions does polarization of this type arise, and can it be prevented through mild interventions?

1.1 Our Models and Results

We address this question in a supervised learning-theoretic framework. Each agent is presented with input points that describe circumstances. And each input point is associated with a label, which represents the appropriate action or opinion. Through experience and observation of examples (input points and their labels), the agent learns a hypothesis, which is a function from the input space $\mathcal{X}$ (which includes previously unseen input points) to the set of labels $\mathcal{Y}$.

To develop intuition, suppose first that two agents see training sets that consist of examples drawn from the same distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and that this distribution is *realizable*; i.e., there is a hypothesis f^* such that $f^*(x) = y$ for every example (x, y) in the support of $\mathcal{D}$. In this classic setting polarization will not arise: given enough examples, both agents will eventually learn hypotheses that almost entirely agree with f^*, and hence with each other.

[1] For more extensive experiments showing such a confirmatory bias, and a model of bounded rationality that explains increasing polarization after exposure to identical information, see the work of Fryer et al. [5].

We introduce and analyze two models that deviate from this basic setting in fundamentally different ways. In our first model — the *Subjective Mixture Model*, presented in Section 3 — agents see different, yet highly overlapping, labels; in this sense the agents each have their own 'subjective' views based on differing personal histories and perspectives.

To formalize this, suppose that two agents are associated with two realizable distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathcal{X} \times \mathcal{Y}$, which have the same marginal over $\mathcal{X}$. Learning from these two different distributions can naturally give rise to highly polarized hypotheses. Now suppose that, perhaps with the aim of finding common ground, the two agents actually communicate and share their sources of information, leading them both to observe biased mixtures of the two distributions. Specifically, we assume agent 1 trains on examples drawn from $(1/2 + \epsilon)\mathcal{D}_1 + (1/2 - \epsilon)\mathcal{D}_2$, while agent 2 trains on examples from $(1/2 - \epsilon)\mathcal{D}_1 + (1/2 + \epsilon)\mathcal{D}_2$, for a small $\epsilon > 0$. We show that, even when mixing to almost even proportions, the two agents will still learn dramatically different hypotheses. This result suggests that when labels are subjective, polarization is inevitable.

In our second model — the *Objective Cost Model*, presented in Section 4 — we go back to the 'objective' setting where agents completely agree on labels. The twist is that we introduce an explicit cost for each function, which intuitively depends on its complexity. These cost functions are in part motivated by a number of studies on the capacity of people for processing information [16, 19, 22]. For example, Miller's Law [16] asserts that the average person only holds about 7 features in their working memory, and a recent study of the interpretability of machine learning models found evidence that the average person finds it as difficult to simulate hypotheses of 8 or more features as to simulate an opaque 'black box' [19].

The Objective Cost Model gives rise to instances, and distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, such that having a non-trivial cost of complexity leads agents to learn hypotheses that result in substantial disagreement, even as the number of observed examples goes to infinity. The basic structure of such instances is that they have many dimensions or variables that actually matter, more or less equally, and need to be accounted for to recover the common target hypothesis $f : \mathcal{X} \to \mathcal{Y}$. Any cost that results in simplifying the hypothesis that is used leads to a selection of some dimensions that are paid attention to; slight differences in the training set lead to a different selections of variables.

We emphasize that instances that exhibit the foregoing structure are not artificial constructs, but actually appear in practice. A striking example is the American Housing Survey dataset considered by Mullainathan and Spiess [18, Figure 2]. It illustrates this phenomenon precisely in the context of a LASSO selection of variables, which is shown to vary dramatically based on a random sampling of the data.

Our main result in the Objective Cost Model is that any instance that leads to polarization can be altered by slightly perturbing the underlying distribution, such that if agents observe large enough training sets then with high probability their learned hypotheses will differ only slightly. This suggests that adding some particular forms of bias in information selection can explain, or lead to, consensus. As we discuss in Section 4.2, not any perturbation will do and the bias needs to be introduced judiciously.

Together our results on the Subjective Mixture Model and the Objective Cost Model highlight a big difference between subjective and objective reasons for polarization: differences due to subjective opinions can be very difficult to overcome, while differences due to objective cost can be corrected with slight, but particular, bias.

1.2 Related Work

Our approach to modeling beliefs and disagreement is most closely related to, but fundamentally different from, that of the social learning literature [6, 7, 17, 24]. This line of work is built on a paradigm by which there is some uncertainty about a state of the world that affects the payoffs to

various potential decisions; it makes heavy use of Bayes' rule, and boundedly-rational variations thereof (e.g., [5, 29]).

As a representative example, consider the classic model of Smith and Sørensen [24]. In this model, the world has a discrete set of payoff-relevant states. Agents observe private signals about the state of the world, and compute (via Bayes' rule) their private beliefs. Moreover, each agent has a private type. Finally, there is a utility function that depends on the agent's type, the state of the world, and the action taken by the agent. To decide how to act, agents again apply Bayes' rule to their own types, their private beliefs, and the history of actions taken by previous agents. Under some parameter constellations, polarization can occur, in the sense that agents converge to multiple limit beliefs.

The paper of Smith and Sørensen [24], and the social learning literature more generally, differ from our work in several ways. First, in the social learning literature 'beliefs' are about the state of the world (it might have been more appropriately named the 'social updating' literature), whereas in our work beliefs are manifested in richer models that people use to best describe the world around them. The fact that having different beliefs corresponds to differences in models captures the fact that people genuinely have individual ways of viewing the world and pay attention to various issues, rather than simply disagreeing about a particular probability. Second, our Objective Cost Model provides a natural way to quantify complexity and costs. More discriminating hypotheses or models, which pay attention to more information or dimensions and make finer and more discriminating predictions, are more complex and costly than ones that pay attention to less information and/or make less discriminating predictions. This differs from the type of cost used in the Bayesian framework, which more naturally relates to the information content of input signals rather than of processing the information.[2] Third, arguably, our viewpoint of machine learning theory is a closer fit with how companies and other organizations operate. They generally are not doing Bayesian estimation, but instead use past data to develop methods of predicting outcomes — often explicitly using machine learning techniques.

Much further afield, there is a rich literature on stochastic models of opinion dynamics, most notably that masterpiece of creative nomenclature, the *Voter Model* [11, 21]. In this model, each node in a graph has one of two colors (say red or blue). At each step a random node is selected, and it adopts the color of one of its neighbors, also selected at random. Although the classic model inevitably leads to consensus, variants thereof have been proposed to capture polarization [27].

2 PRELIMINARIES

Throughout this paper, we work with an input space $\mathcal{X}$ and set of labels $\mathcal{Y} = \{-1, +1\}$.

2.1 Distributions and Distances

We consider distributions $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and denote by $\mathcal{D}{\downarrow}\mathcal{X}$ the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$.

For any two distributions $\mathcal{P}$ and $\mathcal{P}'$ over a domain $\mathcal{X}$, we denote their *total variation* distance by

$$\mathrm{TV}(\mathcal{P}, \mathcal{P}') := \sup_{X \subseteq \mathcal{X}} |\mathcal{P}(X) - \mathcal{P}'(X)|.$$

[2] An exception is the rational inattention literature (e.g., [23]), but that still deals with how many signals are input, not with the complexity of processing those signals. Closer to our motivation is the work of Rubinstein [20], who explicitly models decision making and its costs, but in a different paradigm.

For ease of exposition, we define the L_1 distance between these distributions by $\|\mathcal{P} - \mathcal{P}'\|_1 = 2\text{TV}(\mathcal{P}, \mathcal{P}')$. Note that when $\mathcal{X}$ is a finite domain and when p_x and p'_x respectively denote distributions $\mathcal{P}$ and $\mathcal{P}'$ on $x \in \mathcal{X}$, we have

$$\|\mathcal{P} - \mathcal{P}'\|_1 = \sum_{x \in \mathcal{X}} |p_x - p'_x| = 2 \sum_{x : p_x \geq p'_x} (p_x - p'_x) = 2 \sum_{x : p_x < p'_x} (p'_x - p_x).$$

For any two distributions $\mathcal{D}$ and $\mathcal{D}'$ over $\mathcal{X} \times \mathcal{Y}$, we say that $\mathcal{D}'$ *matches the conditional label distributions of* $\mathcal{D}$ if for all $x \in \mathcal{X}$,[3]

$$\Pr_{(x,y)\sim\mathcal{D}}[y \mid x] = \Pr_{(x,y)\sim\mathcal{D}'}[y \mid x].$$

When $\mathcal{D}'$ matches the conditional label distributions of $\mathcal{D}$ we use $\|\mathcal{D} - \mathcal{D}'\|$ and $\|\mathcal{D}{\downarrow}\mathcal{X} - \mathcal{D}'{\downarrow}\mathcal{X}\|$ interchangeably.

2.2 Hypotheses and Error

We also consider a hypothesis class $\mathcal{F}$ such that for all $f \in \mathcal{F}$, $f : \mathcal{X} \to \mathcal{Y}$. For any hypothesis f, the *error* of f on $\mathcal{D}$ is described by

$$\text{err}_{\mathcal{D}}(f) \coloneqq \Pr_{(x,y)\sim\mathcal{D}}[f(x) \neq y].$$

We say that $\mathcal{D}$ is *realizable* if there exists $f \in \mathcal{F}$ such that $\text{err}_{\mathcal{D}}(f) = 0$. Furthermore, for a training set of m labeled input points $S = \{(x^i, y^i)\}_{i\in[m]}$, we denote the *empirical error* by

$$\text{err}_S(f) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left(f(x^i) \neq y^i\right).$$

We are interested in the disagreement between different hypotheses. For any $f, f' \in \mathcal{F}$, we denote the *disagreement* of f and f' on distribution $\mathcal{D}$ by

$$\Delta_{\mathcal{D}}(f, f') \coloneqq \Pr_{x\sim\mathcal{D}{\downarrow}\mathcal{X}}[f(x) \neq f'(x)].$$

For any set of hypotheses $\mathcal{H}$, we define the *diameter* of $\mathcal{H}$, denoted by

$$\text{diam}_{\mathcal{D}}(\mathcal{H}) \coloneqq \max_{f,f'\in\mathcal{H}} \Delta_{\mathcal{D}}(f, f'),$$

as the largest disagreement between two hypotheses in this class. Note that the disagreement between two hypotheses and the diameter of a hypothesis class do not depend on the labels of distribution $\mathcal{D}$. Therefore, with a slight abuse of notation, we use $\mathcal{D}$ in place of $\mathcal{D}{\downarrow}\mathcal{X}$ in these notations, or suppress the distribution in the notation for diameter and disagreement when it is clear from context.

2.3 Sample Complexity

A number of learning-theoretic tools are available for linking the empirical and true error of a hypothesis. An especially well-known measure of statistical complexity of a class of hypotheses is called *VC dimension*. $\mathcal{F}$ is said to have VC dimension $\text{VCD}(\mathcal{F}) = d$ if the largest set of input points $X \subseteq \mathcal{X}$ for which $\{(f(x))_{x\in X} \mid f \in \mathcal{F}\} = 2^X$ has cardinality $|X| = d$. For any class of hypotheses $\mathcal{F}$, $\epsilon > 0$, and $\delta > 0$, there is

$$m_{\epsilon,\delta} \in O\left(\frac{1}{\epsilon^2}\left(\text{VCD}(\mathcal{F}) + \ln\left(\frac{1}{\delta}\right)\right)\right) \tag{1}$$

[3]This definition can be extended to hold for all $x \in X$ when $X \subseteq \mathcal{X}$ includes all but a measure zero (under $\mathcal{D}$ and $\mathcal{D}'$) subset of $\mathcal{X}$.

such that for any distribution $\mathcal{D}$ and with probability $1-\delta$ over the choice of set S of at least $m_{\epsilon,\delta}$ i.i.d. samples, for all $f \in \mathcal{F}$, we have $|\mathrm{err}_{\mathcal{D}}(f) - \mathrm{err}_S(f)| \leq \epsilon$.

When $\mathcal{D}$ satisfies certain properties, one may be able to learn the optimal hypothesis using fewer samples than presented in Equation (1). For example, when $\mathcal{D}$ is realizable,

$$m^r_{\epsilon,\delta} \in O\left(\frac{1}{\epsilon}\left(\mathrm{VCD}(\mathcal{F})\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right) \tag{2}$$

is sufficient so that with probability $1-\delta$, any hypothesis f with empirical error $\mathrm{err}_S(f) = 0$ satisfies $\mathrm{err}_{\mathcal{D}}(f) \leq \epsilon$.

Even for distributions that are non-realizable, one can obtain faster learning rates if the Bayes optimal classifier is in $\mathcal{F}$. More formally, if $f^{\mathrm{Bayes}}(x) \coloneqq \mathrm{argmax}_y \Pr[y|x] \in \mathcal{F}$ and for all $x \in \mathcal{X}$, $|\Pr[y|x] - \Pr[-y|x]| \geq \beta$,[4] then

$$m^{\beta}_{\epsilon,\delta} \in O\left(\frac{1}{\beta\epsilon}\left(\mathrm{VCD}(\mathcal{F})\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right) \tag{3}$$

samples are sufficient so that with probability $1-\delta$ the empirical error minimizer $\bar{f} \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{err}_S(f)$ also satisfies

$$\mathrm{err}_{\mathcal{D}}(\bar{f}) - \mathrm{err}_{\mathcal{D}}(f^{\mathrm{Bayes}}) \leq \epsilon.$$

3 THE SUBJECTIVE MIXTURE MODEL

In our first model, the *Subjective Mixture Model*, we represent two world views through two realizable distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $\mathcal{X}\times\mathcal{Y}$, which are consistent with hypotheses f_1 and f_2, that is, $\mathrm{err}_{\mathcal{D}_1}(f_1) = \mathrm{err}_{\mathcal{D}_2}(f_2) = 0$. We consider two agents that, perhaps through communication, end up observing examples from almost identical mixtures of these two distributions, and ask whether they could learn substantially different hypotheses.

Specifically, let $\widetilde{\mathcal{D}}_1 \coloneqq (1-\alpha)\mathcal{D}_1 + \alpha\mathcal{D}_2$ and $\widetilde{\mathcal{D}}_2 \coloneqq (1-\alpha)\mathcal{D}_2 + \alpha\mathcal{D}_1$, for $\alpha < 1/2$. Is it possible that $\alpha = 0.49999$ and two agents learning from $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$, which are almost identical, would reach very different conclusions? The following theorem answers this question in the affirmative.

Theorem 3.1. *Let distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $\mathcal{X}\times\mathcal{Y}$ be two realizable distributions over $\mathcal{F}$ with the same marginal distribution. Let $\widetilde{\mathcal{D}}_1 \coloneqq (1-\alpha)\mathcal{D}_1 + \alpha\mathcal{D}_2$ and $\widetilde{\mathcal{D}}_2 \coloneqq (1-\alpha)\mathcal{D}_2 + \alpha\mathcal{D}_1$, for $\alpha < 1/2$. Then there is*

$$m \in O\left(\frac{1}{\left(\frac{1}{2}-\alpha\right)^2\epsilon}\left(\mathrm{VCD}(\mathcal{F})\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

such that if sample sets S_1, S_2, $\widetilde{S}_1$, $\widetilde{S}_2$ of size at least m are sampled from $\mathcal{D}_1$, $\mathcal{D}_2$, $\widetilde{\mathcal{D}}_1$, and $\widetilde{\mathcal{D}}_2$, respectively, then with probability at least $1-\delta$

$$\Delta\left(\widetilde{f}_1, \widetilde{f}_2\right) \geq \Delta(f_1, f_2) - \epsilon,$$

where $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{err}_{S_i}(f)$ and $\widetilde{f}_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{err}_{\widetilde{S}_i}(f)$ for $i \in \{1,2\}$.

In particular, consider two agents who start with world views $\mathcal{D}_1$ and $\mathcal{D}_2$ that are consistent with two hypotheses f_1 and f_2 with large disagreement $\Delta_{\mathcal{D}}(f_1, f_2)$ (where $\mathcal{D}$ is the marginal distribution on inputs agreeing with the marginals $\mathcal{D}_1{\downarrow}\mathcal{X} = \mathcal{D}_2{\downarrow}\mathcal{X}$). Even if each agent attempts to see the world from the other's perspective, they would observe distributions $\widetilde{\mathcal{D}}_1 \coloneqq (1-\alpha)\mathcal{D}_1 + \alpha\mathcal{D}_2$ and

[4] This property is known as the *Massart condition*. The statistical and computational aspects of distributions satisfying it have long been of interest [1–3, 13].

$\widetilde{\mathcal{D}}_2 \coloneqq (1-\alpha)\mathcal{D}_2 + \alpha\mathcal{D}_1$, and still arrive at two hypotheses $\widetilde{f_1}$ and $\widetilde{f_2}$ that are almost as polarized as before.

At a high level, the reason for this phenomenon is that hypotheses f_1 and f_2 remain the optimal hypotheses for distributions $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$. This is due to the fact that for any $(x, y) \sim \mathcal{D}_i$, (x, y) appears with higher probability than $(x, -y)$ in $\widetilde{\mathcal{D}}_i$. So, the optimal classifier for $\widetilde{\mathcal{D}}$ should label x just as the perfect hypothesis would label it on $\mathcal{D}_i$. We formalize this below.

Lemma 3.2. *Let distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $\mathcal{X} \times \mathcal{Y}$ be two realizable distributions with the same marginal distribution. Then,*

$$\operatorname*{argmin}_{f \in \mathcal{F}} \operatorname{err}_{\widetilde{\mathcal{D}}_i}(f) = \left\{ f \in \mathcal{F} \mid \operatorname{err}_{\mathcal{D}_i}(f) = 0 \right\}, \ \forall i \in \{1, 2\}.$$

Proof. In short, for $\alpha < 1/2$, f_i is the Bayes optimal classifier for $\widetilde{\mathcal{D}}_i$. We provide a detailed proof below, which shows that the disagreement of two hypotheses and their relative errors are closely related on $\widetilde{\mathcal{D}}_i$. Specifically, we will prove that for any classifier f, any f_1 such that $\operatorname{err}_{\mathcal{D}_1}(f_1) = 0$, and for any $\alpha < \frac{1}{2}$,

$$(1-2\alpha)\Delta_{\widetilde{\mathcal{D}}_1}(f, f_1) \le \operatorname{err}_{\widetilde{\mathcal{D}}_1}(f) - \operatorname{err}_{\widetilde{\mathcal{D}}_1}(f_1) \le \Delta_{\widetilde{\mathcal{D}}_1}(f, f_1), \tag{4}$$

and similarly for $\widetilde{\mathcal{D}}_2$. Note that Equation (4) proves the claim directly, as hypothesis f minimizes $\operatorname{err}_{\widetilde{\mathcal{D}}_i}$ if and only if $\Delta_{\widetilde{\mathcal{D}}_i}(f, f_i) = 0$, which implies that $\operatorname{err}_{\mathcal{D}_i}(f) = \operatorname{err}_{\mathcal{D}_i}(f_i) = 0$.

The rightmost inequality in Equation (4) holds because for any $(x, y) \sim \widetilde{\mathcal{D}}_1$ such that $f(x) \neq f_1(x)$, either f is incorrect, in which case $\mathbb{I}(f(x) \neq y)) - \mathbb{I}(f_1(x) \neq y)) = 1$ or f_1 is incorrect, in which case $\mathbb{I}(f(x) \neq y)) - \mathbb{I}(f_1(x) \neq y)) = -1$. In both cases, $\mathbb{I}(f(x) \neq y)) - \mathbb{I}(f_1(x) \neq y)) \le \mathbb{I}(f(x) \neq f_1(x))$.

For the leftmost inequality, let $A \coloneqq \{x \mid f(x) \neq f_1(x)\}$ be the region of disagreement between f and f_1. Note that

$$\operatorname{err}_{\widetilde{\mathcal{D}}_1}(f) - \operatorname{err}_{\widetilde{\mathcal{D}}_1}(f_1) = \Pr[A] \left(\operatorname{err}_{\widetilde{\mathcal{D}}_1|A}(f) - \operatorname{err}_{\widetilde{\mathcal{D}}_1|A}(f_1) \right),$$

since f_1 and f incur the same error outside of A. On the other hand, exactly one of f and f_1 makes a mistake on any $(x, y) \sim \widetilde{\mathcal{D}}_1|A$. Therefore, $\operatorname{err}_{\widetilde{\mathcal{D}}_1|A}(f) + \operatorname{err}_{\widetilde{\mathcal{D}}_1|A}(f_1) = 1$. So,

$$\operatorname{err}_{\widetilde{\mathcal{D}}_1}(f) - \operatorname{err}_{\widetilde{\mathcal{D}}_1}(f_1) = \Pr[A] \left(1 - 2 \operatorname{err}_{\widetilde{\mathcal{D}}_1|A}(f_1) \right).$$

By the definition of $\widetilde{\mathcal{D}}_1 \coloneqq (1-\alpha)\mathcal{D}_1 + \alpha\mathcal{D}_2$, we have

$$\operatorname{err}_{\widetilde{\mathcal{D}}_1|A}(f_1) = (1-\alpha) \operatorname{err}_{\mathcal{D}_1|A}(f_1) + \alpha \operatorname{err}_{\mathcal{D}_2|A}(f_1) \le \alpha,$$

where the last transition is due to the fact that f_1 is the perfect classifier for $\mathcal{D}_1$. Using the fact that $\Pr[A] = \Delta_{\widetilde{\mathcal{D}}_1}(f, f_1)$, we have $(1-2\alpha)\Delta_{\widetilde{\mathcal{D}}_1}(f, f_1) \le \operatorname{err}_{\widetilde{\mathcal{D}}_1}(f) - \operatorname{err}_{\widetilde{\mathcal{D}}_1}(f_1)$. ∎

Proof of Theorem 3.1. First note that regardless of the choice of $f_i^* \in \operatorname{argmin}_{f \in \mathcal{F}} \operatorname{err}_{\mathcal{D}_i}(f)$, $\Delta(f_1^*, f_2^*)$ is the same. This is simply because for any $f_i^*, f_i' \in \operatorname{argmin}_{f \in \mathcal{F}} \operatorname{err}_{\mathcal{D}_i}(f)$ we have that $\operatorname{err}_{\mathcal{D}_i}(f_i^*) = \operatorname{err}_{\mathcal{D}_i}(f_i') = 0$. So, these two hypotheses agree on the label of every input point, except for a measure zero subset of the domain, which does not affect error or disagreement. Therefore, $\Delta(f_1^*, f_2^*) = \Delta(f_1', f_2')$. We let $\Delta \coloneqq \Delta(f_1^*, f_2^*)$ represent this fixed quantity. Note that by the sample complexity of realizable learning given in Equation (2), with probability $1 - \delta/2$,

$$\Delta\left(f_1, f_1^*\right) = \operatorname{err}_{\mathcal{D}_1}(f_1) \le \frac{\epsilon}{4} \quad \text{and} \quad \Delta\left(f_2, f_2^*\right) = \operatorname{err}_{\mathcal{D}_2}(f_2) \le \frac{\epsilon}{4},$$

where $f_i = \text{argmin}_{f \in \mathcal{F}} \text{err}_{S_i}(f)$. Hence,

$$\Delta(f_1, f_2) \in \left[\Delta - \frac{\epsilon}{2}, \Delta + \frac{\epsilon}{2}\right]. \tag{5}$$

By Lemma 3.2, the sets of optimal hypotheses on $\mathcal{D}_i$ and $\widetilde{\mathcal{D}}_i$ are the same. Therefore, regardless of the choice of $\widetilde{f}_i^* \in \text{argmin}_{f \in \mathcal{F}} \text{err}_{\widetilde{\mathcal{D}}_i}(f)$, we have that

$$\Delta = \Delta(\widetilde{f}_1^*, \widetilde{f}_2^*). \tag{6}$$

Note that, for $i \in \{1, 2\}$, distribution $\widetilde{\mathcal{D}}_i$ is such that the Bayes optimal hypothesis, which is equivalent to $\widetilde{f}_i$, belongs to $\mathcal{F}$. Using the sample complexity result of Equation (3) for such distributions, we have that with probability $1 - \delta/2$,

$$\text{err}_{\widetilde{\mathcal{D}}_i}(\widetilde{f}_i) - \text{err}_{\widetilde{\mathcal{D}}_i}(\widetilde{f}_i^*) \leq \frac{\epsilon(1 - 2\alpha)}{4} \quad \text{for } i \in \{1, 2\},$$

where $\widetilde{f}_i = \text{argmin}_{f \in \mathcal{F}} \text{err}_{\widetilde{S}_i}(f)$. Using Equation (4), it holds that

$$\Delta\left(\widetilde{f}_i, \widetilde{f}_i^*\right) \leq \frac{1}{1 - 2\alpha}\left(\text{err}_{\widetilde{\mathcal{D}}_i}(\widetilde{f}_i) - \text{err}_{\widetilde{\mathcal{D}}_i}(\widetilde{f}_i^*)\right) \leq \frac{\epsilon}{4} \quad \text{for } i \in \{1, 2\}.$$

By Equation (6), $\Delta(\widetilde{f}_1, \widetilde{f}_2) \in [\Delta - \frac{\epsilon}{2}, \Delta + \frac{\epsilon}{2}]$. Using Equation (5), we conclude that $\Delta(\widetilde{f}_1, \widetilde{f}_2) \geq \Delta(f_1, f_2) - \epsilon$. ■

The results above apply to situations in which the two agents see different labelings for the same inputs. Even if the distributions they see are nearly the same, slight differences in the frequencies of labels that they end up observing for the same inputs are enough to allow them to reach very different conclusions in terms of the error-minimizing hypotheses that they adopt. It is in this sense that even people who strive to communicate and find common ground with others can form polarized opinions. This gives one possible explanation for polarization, predicated upon (even tiny) differences in experiences. We next explore a different explanation.

4 THE OBJECTIVE COST MODEL

In this section, we turn our attention to the *Objective Cost Model*, where the common world view is represented by a realizable distribution $\mathcal{D}$ that is consistent with some hypothesis $f^* \in \mathcal{F}$. In this case, any agent who observes a sufficiently large number of labeled samples from $\mathcal{D}$, and adopts the hypothesis that minimizes error on these samples, learns a belief that is in almost full agreement with f^*. Therefore, all error-minimizing agents will arrive at hypotheses that are almost in full agreement with each other.

However, when the error-minimizing hypotheses are very *complex*, the learner may face difficulty learning or interpreting them. Therefore, rather than considering error-minimizing agents, we consider agents who attempt to find a hypothesis that strikes a balance between accuracy and complexity. We first show that when there is a complexity associated with learning functions, it is entirely possible that two agents receiving two i.i.d. training sets from $\mathcal{D}$ would learn hypotheses f_1 and f_2 that are in large disagreement. We then discuss how this can be dealt with by changing the original distribution $\mathcal{D}$ slightly to help the two agents arrive at hypotheses that are mostly in agreement.

4.1 Setup and Initial Observations

Before we proceed, we require some new notation. We assume that $\mathcal{F}$ is accompanied by a known complexity function $\phi(\cdot)$, such that for any $f \in \mathcal{F}$, $\phi(f)$ determines how complex hypothesis f is. Some examples of such complexity measures include monotonic functions of the number of

features in a boolean function, or the depth of a decision list. Additionally, for a distribution $\mathcal{D}$ and a training set S, we denote the cost incurred by any hypothesis through its error and complexity as

$$\text{cost}_{\mathcal{D}}(f) \coloneqq \text{err}_{\mathcal{D}}(f) + \phi(f) \text{ and } \text{cost}_{S}(f) \coloneqq \text{err}_{S}(f) + \phi(f).$$

We start by demonstrating the phenomenon described above in the most extreme case, where there are $f_1, f_2 \in \text{argmin}_{f \in \mathcal{F}} \text{cost}_{\mathcal{D}}(f)$ that have a large disagreement with each other.

Example 4.1. Let $\mathcal{X} \coloneqq \mathbb{R}^d$ and let $\mathcal{F}$ be the class of homogeneous linear separators, i.e.,

$$\mathcal{F} \coloneqq \{f_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x})\}_{\mathbf{w} \in \mathbb{R}^d},$$

where $\text{sign}(z) = +1$ for $z \geq 0$ and -1 otherwise. Let the complexity function be $\phi(f_{\mathbf{w}}) \coloneqq \lambda \|\mathbf{w}\|_0$ for $\lambda \coloneqq \frac{1}{2d}$, i.e., the cost of complexity is λ for each nonzero entry in $\mathbf{w}$. Let $\mathcal{D}$ have weight $\frac{1}{2d}$ on each of $(\mathbf{e}_i, +1)$ and $(-\mathbf{e}_i, -1)$ for all $i \in [d]$, where $\mathbf{e}_i$ is the i'th unit vector. We show that

$$\left\{\mathbf{w} \mid w_i \in \{0, 1\} \text{ for all } i \in [d]\right\} \subseteq \underset{\mathbf{w} \in \mathbb{R}^d}{\text{argmin}} \, \text{cost}_{\mathcal{D}}(f_{\mathbf{w}}). \tag{7}$$

Note that in this case, $f_{\mathbf{0}}$ and $f_{\mathbf{1}}$ (corresponding to the all-zeros and all-ones vectors) disagree on all $(-\mathbf{e}_i, -1)$ for $i \in [d]$, so, $\Delta(f_{\mathbf{0}}, f_{\mathbf{1}}) = \frac{1}{2}$.

Let us show, then, that Equation (7) holds. Note that for any $\mathbf{w} \in \mathbb{R}^d$, rounding $w_i > 0$ to 1 and $w_i < 0$ to -1 does not change its error or cost. Furthermore, for any i, if $w_i = 1$ then $f_{\mathbf{w}}$ labels both $(\mathbf{e}_i, +1)$ and $(-\mathbf{e}_i, -1)$ correctly, if $w_i = 0$ then $f_{\mathbf{w}}$ labels $(\mathbf{e}_i, +1)$ correctly and $(-\mathbf{e}_i, -1)$ incorrectly, and if $w_i = -1$ then $f_{\mathbf{w}}$ labels both $(\mathbf{e}_i, +1)$ and $(-\mathbf{e}_i, -1)$ incorrectly. That is, for any i, setting $w_i = 1$ instead of $w_i = 0$, decreases the error by exactly $\frac{1}{2d}$ and increases the complexity by $\frac{1}{2d}$. Therefore, for all $\mathbf{w} \in \{0, 1\}^d$, $\text{cost}_{\mathcal{D}}(f_{\mathbf{w}}) = \frac{1}{2}$, which is the minimum cost of any $f_{\mathbf{w}} \in \mathcal{F}$.

We can easily leverage this example to demonstrate that there are situations where it is likely that two agents would learn hypotheses that disagree substantially.

Theorem 4.2. *There is a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is realizable for a hypothesis class $\mathcal{F}$, such that for any m and two sets of m i.i.d. samples, S_1 and S_2, with probability at least $\frac{1}{4}$, there are $f_i \in \text{argmin}_{f \in \mathcal{F}} \text{cost}_{S_i}(f)$ for $i \in \{1, 2\}$ such that $\Delta_{\mathcal{D}}(f_1, f_2) > \frac{1}{6}$.*

Note that using specific $f_i \in \text{argmin}_{f \in \mathcal{F}} \text{cost}_{S_i}(f)$ allows us to break ties in a way that is convenient for us. This is purely for ease of exposition: we could formulate the theorem for any cost-minimizing f_i, using slightly messier probability calculations.

Proof of Theorem 4.2. We use the setting of Example 4.1. Note that if S_1 has *more than* a $\frac{1}{2d}$ fraction of its samples on $(-\mathbf{e}_j, -1)$, then there is a cost-minimizing f_1 that has $w_j = 1$, and $w_j = 0$ otherwise. Similarly, if S_2 has *at least* a $\frac{1}{2d}$ fraction of its samples on $(-\mathbf{e}_j, -1)$, then there is a cost-minimizing f_2 that has $w_j = 1$, and $w_j = 0$ otherwise. Using well-known properties of binomial distributions [10], it follows that the probability that $f_1(-\mathbf{e}_j) \neq f_2(-\mathbf{e}_j)$ is at least $1/2$.

Let Z be the random variable whose value is the number of coordinates $j \in [d]$ such that $f_1(-\mathbf{e}_j) \neq f_2(-\mathbf{e}_j)$. It holds that $Z \leq d$, and $\mathbb{E}[Z] \geq d/2$. Therefore, it must be the case that $\Pr[Z \leq d/3] \leq 3/4$, that is, $\Pr[Z > d/3] > 1/4$. Hence, with probability at least $1/4$, $\Delta_{\mathcal{D}}(f_1, f_2) > 1/6$. ∎

While the above example is artificially constructed to demonstrate how objective cost can give rise to polarization, it is worth noting that high levels of disagreement between learned models have also been observed in practice, as we noted in Section 1.1. A possible explanation for why polarization occurs even under 'natural' distributions is that perturbing the polarizing distributions by some unbiased noise does not alleviate the problem.[5] That is, adding unbiased (but realizable)

[5]This point of view is consistent with that of *Smoothed Analysis* [25], which is built on the idea that 'natural' instances can be described by worst-case instances that are then smoothed by nature using unbiased noise.

noise to the above example, such as adding a uniform distribution over the domain that is labeled by the optimal hypothesis, still leads to the same outcome. This is best observed by noting that distribution $\mathcal{D}$ of Example 4.1 is uniform over its support, so noise that does not introduce bias towards specific subsets of the support does not change the distribution at all. Our main result, below, builds on this intuition.

4.2 Introducing Bias to Prevent Polarization

Next, we show that for any problematic distribution $\mathcal{D}$ as in Theorem 4.2, one can carefully design $\widetilde{\mathcal{D}}$ that is close to $\mathcal{D}$ in its marginal distribution and is still realizable with respect to $\mathcal{F}$, such that $\widetilde{\mathcal{D}}$ does not suffer from the same problem. That is, it is always possible to prevent polarization (of the type studied here) by introducing a slight bias into the information selection process. Formally, we prove the following theorem.

Theorem 4.3 (Main). *Consider a hypothesis class $\mathcal{F}$, a realizable distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, an $\alpha \in [0,1]$, and a maximum level of disagreement $\gamma > 0$. Then, there is a realizable distribution $\widetilde{\mathcal{D}}$, such that $\|\widetilde{\mathcal{D}} - \mathcal{D}\|_1 \leq \alpha$, and there is*

$$m \in O\left(\gamma^{-4}\alpha^{-2}\left(\mathrm{VCD}(\mathcal{F}) + \ln\left(\frac{1}{\delta}\right)\right)\right),$$

such that if two sets S_1 and S_2 of size at least m are sampled from $\widetilde{\mathcal{D}}$, then with probability at least $1-\delta$ any two cost-minimizing hypotheses $f_i \in \operatorname{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ for $i \in \{1,2\}$

(1) have at most γ disagreement over $\mathcal{D}$, i.e., $\Delta_{\mathcal{D}}(\widetilde{f}_1, \widetilde{f}_2) \leq \gamma$, and

(2) have a cost that is optimal up to 3α on $\mathcal{D}$, i.e.,

$$\mathrm{cost}_{\mathcal{D}}(\widetilde{f}_i) \leq \operatorname*{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{\mathcal{D}}(f) + 3\alpha.$$

The rest of this subsection is devoted to the theorem's proof. First, to assist with examining disagreement between cost-minimizing hypotheses, we introduce some additional notation. For any ϵ and $\mathcal{D}$, let

$$\mathcal{F}_\epsilon^{\mathcal{D}} := \left\{ f \in \mathcal{F} \;\middle|\; \mathrm{cost}_{\mathcal{D}}(f) \leq \operatorname*{argmin}_{f'\in\mathcal{F}} \mathrm{cost}_{\mathcal{D}}(f') + \epsilon \right\}.$$

This definition is loosely related to the concept of *Rashomon sets* from statistics [4].

At a high level, if $\mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}}) > \Delta$ is large, then unless a large number of samples is used (of the order of $\mathrm{VCD}(\mathcal{F})\epsilon^{-2}$), two learners with samples drawn from $\mathcal{D}$ can learn two hypotheses with disagreement Δ. We will formalize this further later on. But presently let us show that if $\mathrm{diam}(\mathcal{F}_\epsilon^{\mathcal{D}})$ is large, then there is a distribution $\widetilde{\mathcal{D}}$ close to $\mathcal{D}$ with a much smaller $\mathrm{diam}(\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}})$. This implies that learning over the distribution $\widetilde{\mathcal{D}}$ aligns the learning process of two learners and leads to models that have small disagreement.

Lemma 4.4. *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ that is realizable with respect to $\mathcal{F}$, i.e., there is $f^* \in \mathcal{F}$ such that $\mathrm{err}_{\mathcal{D}}(f^*) = 0$. Assume that for some $\epsilon > 0$, $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\mathcal{D}}\right) \geq 2R$ for some $R \in [0, 0.5]$. Then there is a distribution $\mathcal{P}$ with $\mathrm{err}_{\mathcal{P}}(f^*) = 0$, such that for any $\alpha > 0$ and $\widetilde{\mathcal{D}} = (1-\alpha)\mathcal{D} + \alpha\mathcal{P}$, $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}\right) \leq \frac{8\epsilon}{\alpha R}$.*

Proof. Let $\widetilde{f} \in \operatorname{argmax}_{f\in\mathcal{F}_\epsilon^{\mathcal{D}}} \mathrm{err}_{\mathcal{D}}(f)$ be a hypothesis with maximum error in $\mathcal{F}_\epsilon^{\mathcal{D}}$. Let

$$\mathcal{P} := \mathcal{D}|\left\{ x \;\middle|\; \widetilde{f}(x) = f^*(x) \right\}$$

be distribution $\mathcal{D}$ conditioned on the set of input points where $\widetilde{f}$ is correct, and $\widetilde{\mathcal{D}} = (1-\alpha)\mathcal{D} + \alpha\mathcal{P}$. We will show that for all $f \in \mathcal{F}$,

$$\Delta_{\mathcal{D}}(f, \widetilde{f}) > \frac{4\epsilon}{\alpha R} \Longrightarrow f \notin \mathcal{F}_{\epsilon}^{\widetilde{\mathcal{D}}}. \tag{8}$$

Note that this directly implies the lemma, using the triangle inequality.

Before we prove this claim, let us prepare to analyze the change in cost caused by changing $\mathcal{D}$ to $\widetilde{\mathcal{D}}$. Fix $f \in \mathcal{F}$, and let

$$\begin{aligned} \Delta &\coloneqq \left(\mathrm{cost}_{\widetilde{\mathcal{D}}}(\widetilde{f}) - \mathrm{cost}_{\widetilde{\mathcal{D}}}(f)\right) - \left(\mathrm{cost}_{\mathcal{D}}(\widetilde{f}) - \mathrm{cost}_{\mathcal{D}}(f)\right) \\ &= -\alpha\left(\mathrm{err}_{\mathcal{D}}(\widetilde{f}) - \mathrm{err}_{\mathcal{D}}(f) + \mathrm{err}_{\mathcal{P}}(f)\right). \end{aligned} \tag{9}$$

In addition, define

$$A \coloneqq \left\{x \in \mathcal{X} \;\middle|\; \widetilde{f}(x) = f^*(x) \text{ and } f(x) \neq f^*(x)\right\},$$

and

$$C \coloneqq \left\{x \in \mathcal{X} \;\middle|\; \widetilde{f}(x) \neq f^*(x) \text{ and } f(x) = f^*(x)\right\}.$$

Then it holds that

$$\mathrm{err}_{\mathcal{D}}(\widetilde{f}) - \mathrm{err}_{\mathcal{D}}(f) = \Pr_{x\sim\mathcal{D}}[x \in C] - \Pr_{x\sim\mathcal{D}}[x \in A], \tag{10}$$

and

$$\mathrm{err}_{\mathcal{P}}(f) = \frac{\Pr_{x\sim\mathcal{D}}[x \in A]}{1 - \mathrm{err}_{\mathcal{D}}(\widetilde{f})}. \tag{11}$$

Finally, it holds that

$$\Pr_{x\sim\mathcal{D}}[x \in A] + \Pr_{x\sim\mathcal{D}}[x \in C] = \Delta_{\mathcal{D}}(f, \widetilde{f}), \tag{12}$$

and therefore, using Equation (10),

$$2 \cdot \Pr_{x\sim\mathcal{D}}[x \in C] = \Delta_{\mathcal{D}}(f, \widetilde{f}) + \mathrm{err}_{\mathcal{D}}(\widetilde{f}) - \mathrm{err}_{\mathcal{D}}(f). \tag{13}$$

Using Equations (9), (10), and (11),

$$\begin{aligned} \Delta &= -\alpha\left(\frac{\Pr_{x\sim\mathcal{D}}[x \in A]}{1 - \mathrm{err}_{\mathcal{D}}(\widetilde{f})} + \Pr_{x\sim\mathcal{D}}[x \in C] - \Pr_{x\sim\mathcal{D}}[x \in A]\right) \\ &\leq -\alpha\left(\Pr_{x\sim\mathcal{D}}[x \in A] \cdot (1 + \mathrm{err}_{\mathcal{D}}(\widetilde{f})) + \Pr_{x\sim\mathcal{D}}[x \in C] - \Pr_{x\sim\mathcal{D}}[x \in A]\right) \\ &= -\alpha\left(\Pr_{x\sim\mathcal{D}}[x \in A] \cdot \mathrm{err}_{\mathcal{D}}(\widetilde{f}) + \Pr_{x\sim\mathcal{D}}[x \in C]\right) \qquad (14) \\ &= -\alpha\left(\Pr_{x\sim\mathcal{D}}[x \in A] \cdot \mathrm{err}_{\mathcal{D}}(\widetilde{f}) + \frac{\Delta_{\mathcal{D}}(f, \widetilde{f}) + \mathrm{err}_{\mathcal{D}}(\widetilde{f}) - \mathrm{err}_{\mathcal{D}}(f)}{2}\right), \qquad (15) \end{aligned}$$

where the last transition follows from Equation (13).

We can now complete the proof by establishing Equation (8). Let $f \in \mathcal{F}$ such that $\Delta_{\mathcal{D}}(f, \widetilde{f}) > \frac{4\epsilon}{\alpha R}$.

First consider the case of f for which $\mathrm{err}_{\mathcal{D}}(f) \leq \mathrm{err}_{\mathcal{D}}(\widetilde{f})$. Using Equation (15),

$$\Delta \leq -\frac{\alpha}{2}\Delta_{\mathcal{D}}(f, \widetilde{f}) < -2\epsilon.$$

Since $\widetilde{f} \in \mathcal{F}_{\epsilon}^{\mathcal{D}}$ and $\Delta < -2\epsilon$, we have that $f \notin \mathcal{F}_{\epsilon}^{\widetilde{\mathcal{D}}}$.

Next, consider the case of f for which $\mathrm{err}_{\mathcal{D}}(f) > \mathrm{err}_{\mathcal{D}}(\widetilde{f})$. Therefore, we have that

$$\Pr_{x\sim\mathcal{D}}[x \in A] > \Pr_{x\sim\mathcal{D}}[x \in C].$$

Plugging this into Equation (12), we have that

$$\Pr_{x\sim\mathcal{D}}[x \in A] > \frac{1}{2}\Delta_{\mathcal{D}}(f, \widetilde{f}).$$

Moreover, by the choice of $\widetilde{f}$, $\mathrm{err}_{\mathcal{D}}(\widetilde{f}) \geq R$, otherwise $\mathrm{diam}(\mathcal{F}_\epsilon^D) < 2R$. By Equation (14), we conclude that

$$\Delta \leq -\alpha \cdot \Pr_{x\sim\mathcal{D}}[x \in A] \cdot \mathrm{err}_{\mathcal{D}}(\widetilde{f}) < -2\epsilon. \tag{16}$$

As before, it follows that $f \notin \mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$. ∎

We now return to the learning-theoretic guarantee we advertised earlier.

Lemma 4.5. *For any two distributions $\mathcal{D}$ and $\mathcal{D}'$, assume that $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\mathcal{D}'}\right) \leq \gamma$. Then there is*

$$m \in O\left(\frac{1}{\epsilon^2}\left(\mathrm{VCD}(\mathcal{F}) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

such that for two sample sets S_1 and S_2 of size at least m from distribution $\mathcal{D}'$, with probability $1-\delta$, if $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ for $i \in \{1,2\}$, it holds that $f_1, f_2 \in \mathcal{F}_\epsilon^{\mathcal{D}'}$ and $\Delta_{\mathcal{D}}(f_1, f_2) \leq \gamma$.

Proof. It is sufficient to show that $f_1, f_2 \in \mathcal{F}_\epsilon^{\mathcal{D}'}$. In turn, it suffices to show m is large enough that for a sample set S of size m, the hypothesis f_S that minimizes the empirical cost on S is in the set $\mathcal{F}_\epsilon^{\mathcal{D}'}$ with probability $1-\delta/2$.

Since $\phi(f)$ is fixed (independently of samples), the concentration of $\mathrm{cost}(f)$ only depends on the concentration of the error. Using Equation (1), we have that with probability $1-\delta/2$, for all $f \in \mathcal{F}$, $|\mathrm{cost}_{\mathcal{D}'}(f) - \mathrm{cost}_S(f)| \leq \epsilon/2$. Therefore, for $\widetilde{f}^* \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{\mathcal{D}'}(f)$ and $f_S \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_S(f)$, we have

$$\mathrm{cost}_{\mathcal{D}'}(f_S) \leq \mathrm{cost}_S(f_S) + \frac{\epsilon}{2} \leq \mathrm{cost}_S(\widetilde{f}^*) + \frac{\epsilon}{2} \leq \mathrm{cost}_{\mathcal{D}'}(\widetilde{f}^*) + \epsilon.$$

This proves that $f_S \in \mathcal{F}_\epsilon^{\mathcal{D}'}$ with probability $1-\delta/2$. By the union bound, with probability $1-\delta$, $f_1, f_2 \in \mathcal{F}_\epsilon^{\mathcal{D}'}$. ∎

We are now ready to prove the theorem.

Proof of Theorem 4.3. Let $\epsilon = \alpha\gamma^2/16$. If $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\mathcal{D}}\right) \leq \gamma$, let $\widetilde{\mathcal{D}} := \mathcal{D}$. By Lemma 4.5, $\Delta_{\mathcal{D}}(\widetilde{f}_1, \widetilde{f}_2) \leq \gamma$ and $\widetilde{f}_1, \widetilde{f}_2 \in \mathcal{F}_\epsilon^{\mathcal{D}}$, so $\mathrm{cost}_{\mathcal{D}}(\widetilde{f}_i)$ is within $\epsilon = \alpha\gamma^2/16 \leq \alpha$ of the optimal.

If $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\mathcal{D}}\right) > \gamma$, define $\widetilde{\mathcal{D}}$ as in Lemma 4.4. Then, we have

$$\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}\right) \leq \frac{8\epsilon}{\alpha\gamma/2} = \gamma.$$

By Lemma 4.5, $\Delta_{\mathcal{D}}(\widetilde{f}_1, \widetilde{f}_2) \leq \gamma$ and $\widetilde{f}_1, \widetilde{f}_2 \in \mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$, so for $i \in \{1,2\}$,

$$\mathrm{cost}_{\widetilde{\mathcal{D}}}(\widetilde{f}_i) \leq \operatorname*{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{\widetilde{\mathcal{D}}}(f) + \epsilon.$$

Moreover, by definition of $\widetilde{D}$, for any $f \in \mathcal{F}$ it holds that $\left|\mathrm{cost}_{\mathcal{D}}(f) - \mathrm{cost}_{\widetilde{\mathcal{D}}}(f)\right| \leq \alpha$. Therefore,

$$\mathrm{cost}_{\widetilde{\mathcal{D}}}(\widetilde{f}_i) - \operatorname*{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{\widetilde{\mathcal{D}}}(f) \leq \epsilon + 2\alpha \leq 3\alpha.$$

∎

We remark that the choice of $\widetilde{f} \in \mathcal{F}_\epsilon^{\mathcal{D}}$ with the highest error in Lemma 4.4 biases the agents' learning process towards *less complex* but also *less accurate models*. It is not hard to see that we could have biased the distribution towards any choice of $f \in \mathcal{F}_\epsilon^{\mathcal{D}}$ while getting a guarantee of $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}) \in O(\frac{\epsilon}{\alpha \mathrm{err}_{\mathcal{D}}(f)})$ (See Equation (16)).

One may ask, then, whether it is possible to introduce bias towards the error-zero hypothesis f^* when $f^* \in \mathcal{F}_\epsilon^{\mathcal{D}}$ (even though the bound in the preceding paragraph is useless because $\mathrm{err}_{\mathcal{D}}(f^*) = 0$). The answer is, perhaps surprisingly, no. To see this, take for example a hypothesis class $\mathcal{F} := \{f_X \mid X \subseteq \mathcal{X}\}$ such that $f_X(x) = 1$ if $x \in X$ and $f_X(x) = 0$ otherwise. Consider the complexity function $\phi(f_X) := \frac{|X|}{d}$ and distribution $\mathcal{D}$ that is uniform over $(x, 1)$ for $x \in \mathcal{X}$. Any distribution $\mathcal{D}'$ that is not fully uniform on $\mathcal{X}$ promotes the hypothesis f_X where $x \in X$ only if the distribution has at least $1/d$ weight on $(x, 1)$. On the other hand, the uniform distribution, which is $\mathcal{D}$ itself, yields equal cost for all hypotheses. Therefore, no distribution can promote f^* specifically.

4.3 Lower Bound

As we observed in Theorem 4.3, for any desired maximum level of disagreement between two agents, γ, and any desired level of intervention α, every distribution $\mathcal{D}$ can be changed to a nearby distribution $\mathcal{D}'$ at distance α, such that learners who receive a *large enough* number of samples from $\mathcal{D}'$ have disagreement of less than γ. Theorem 4.3 shows that the number of samples needed for this to work is at most $O\left(\gamma^{-4}\alpha^{-2}\mathrm{VCD}(\mathcal{F})\right)$. We next provide a lower bound that shows that the number of samples needed for the agents to avoid polarization indeed has to increase with $\frac{1}{\alpha}$ and $\frac{1}{\gamma}$. That is, we succeed at having smaller disagreement between agents and making only small change to the distribution only if agents form their hypotheses after having acquired a large number of observations.

Theorem 4.6. *Let $m(\alpha, \gamma, d, \delta)$ be as follows: For any distribution $\mathcal{D}$ on domain $\mathcal{X}$ and any hypothesis class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ with VC dimension d, and any cost function ϕ over $\mathcal{F}$, there exists $\mathcal{D}'$ such that $\|\mathcal{D} - \mathcal{D}'\|_1 \leq \alpha$, $\mathcal{D}$ and $\mathcal{D}'$ have the same conditional label distributions, and for any $m \geq m(\alpha, \gamma, d, \delta)$, with probability $1 - \delta$, $\Delta_{\mathcal{D}}(f_1, d_2) \leq \gamma$, where $f_i \in \mathrm{argmin}_{f \in \mathcal{F}} \mathrm{cost}_{S_i}(f)$ and S_1 and S_2 are two i.i.d. sample sets of size m from $\mathcal{D}'$. Then for any $\alpha < \frac{1}{3}$, $\gamma < \frac{1}{2}$, and $d \geq 1/\gamma$, we have that*

$$m\left(\alpha, \gamma, d, \frac{1}{4}\right) \in \Omega\left(\frac{d}{\alpha^2} \ln\left(\frac{1}{\gamma}\right)\right).$$

We remark that there is a gap in terms of the dependence on parameter $\frac{1}{\gamma}$ between the upper bound (Theorem 4.3) and the lower bound (Theorem 4.6). This is perhaps good news. After all, we may be able to avoid polarization with a significantly smaller number of samples than that prescribed by Theorem 4.3. In Appendix A, we discuss in more detail the source of this gap between the upper and lower bound, and present a possible path forward towards an improved upper bound.

Turning to the proof of Theorem 4.6, we require the Chernoff-type anti-concentration bound stated below [30].

Lemma 4.7 (Anti-concentration). *Let $X_1, \ldots, X_m$ be i.i.d. Bernoulli random variables such that $p(X_i = 1) = p$. Then, for $\alpha < 1/2$ and $p < 1/2$ such that $\alpha^2 pm \geq 3$, we have that*

$$\Pr\left[\sum_{i=1}^{m} X_i > mp(1+\alpha)\right] \geq \exp\left(-9\alpha^2 mp\right),$$

and

$$\Pr\left[\sum_{i=1}^{m} X_i < mp(1-\alpha)\right] \geq \exp\left(-9\alpha^2 mp\right).$$

Proof of Theorem 4.6. Assume for contradiction that $m \leq \frac{c \cdot d}{\alpha^2} \ln(1/\gamma)$ for a constant c that we will set later. We construct a 'hard' example. Specifically, we work with a general domain $\mathcal{X}$ of size d. We also consider a hypothesis class $\mathcal{F} \coloneqq \{f_X \mid X \subseteq \mathcal{X}\}$ such that $f_X(x) = 1$ if $x \in X$ and $f_X(x) = 0$ otherwise. Note that $\text{VCD}(\mathcal{F}) = d$. We use the complexity function $\phi(f_X) \coloneqq \frac{|X|}{d}$, that is, the complexity of each classifier is the fraction of the domain points it labels 1.

To complete the construction, let $\mathcal{D}$ be the uniform distribution on the set of labeled instances $\{(x, +1)\}_{x \in \mathcal{X}}$. Note that this distribution is realizable using function $f_{\mathcal{X}} \in \mathcal{F}$.

Any distribution $\mathcal{D}'$ whose conditional label distributions match that of $\mathcal{D}$ on domain $\mathcal{X}$ can be represented by point masses p_x for all $x \in \mathcal{X}$. Furthermore, for any such $\mathcal{D}'$ that also satisfies $\|\mathcal{D} - \mathcal{D}'\|_1 \leq \alpha$, $\mathcal{X}$ can be divided to two sets

$$\mathcal{X}^+ \coloneqq \{x \in \mathcal{X} \mid p_x \geq 1/d\}$$

and

$$\mathcal{X}^- \coloneqq \{x \in \mathcal{X} \mid p_x < 1/d\}$$

such that

$$\sum_{x \in \mathcal{X}^+} \left(p_x - \frac{1}{d}\right) = \sum_{x \in \mathcal{X}^-} \left(\frac{1}{d} - p_x\right) \leq \frac{\alpha}{2}. \tag{17}$$

Assume without loss of generality that $|\mathcal{X}^+| \geq d/2$; the case of $|\mathcal{X}^-| \geq d/2$ follows similarly. Finally, let

$$A^+ \coloneqq \left\{x \in \mathcal{X}^+ \;\middle|\; p_x - \frac{1}{d} < \frac{2\alpha}{d}\right\}.$$

Consider a sample set $S \sim (\mathcal{D}')^m$ and let $\hat{p}_x$ represent the mass that its empirical distribution assigns to labeled instance $(x, +1)$. For $x \in A^+$, define a random variable $\mathcal{R}_x$, such that $\mathcal{R}_x = 1$ if $\hat{p}_x < \frac{1}{d}$.

Claim 4.8.

$$\mathop{\mathbb{E}}_{S \sim (\mathcal{D}')^m} \left[\sum_{x \in A^+} \mathcal{R}_x\right] \geq 8d\gamma.$$

Proof of Claim. We first show that for any $x \in A^+$, with probability at least 32γ we have $\hat{p}_x < \frac{1}{d}$. To see why this is the case, we use Lemma 4.7. That is, for $x \in A^+$, we have that $p_x \leq \frac{1}{d}(1 + 2\alpha)$, so

$$\Pr\left[\hat{p}_x < \frac{1}{d}\right] \geq \Pr\left[\hat{p}_x < p_x(1 - 2\alpha)\right] \geq \exp\left(-36\alpha^2 p_x m\right) \geq \exp\left(-36c \ln\left(\frac{1}{\gamma}\right)\right) \geq 32\gamma,$$

where constant c is chosen to satisfy the last transition.

Now it suffices to show that $|A^+| > \frac{d}{4}$. To see why this is true, note that, by Equation (17) and $|X^+| \geq d/2$, the average of $p_x - \frac{1}{d}$ over elements of $\mathcal{X}^+$ is at most α/d. Therefore, at least half of these elements, i.e., $|A^+| \geq d/4$ elements overall, must have $p_x - \frac{1}{d} < \frac{2\alpha}{d}$. This establishes the claim. ■

Claim 4.9.

$$\Pr_{S \sim (\mathcal{D}')^m} \left[\sum_{x \in A^+} \mathcal{R}_x \geq 4d\gamma\right] \geq \frac{1}{2}.$$

Proof of Claim. Note that the $\hat{p}_x$ variables are not independent. However, it is well known that they are *negatively associated* [9].[6] Moreover, the $\mathcal{R}_x$ variables are all monotonically decreasing

[6] At a high level, sum of *negatively associated* variables enjoy similar or stronger concentration properties compared to the sum of independent Bernoulli variable.

functions of the $\hat{p}_x$ variables.[7] It is known that such random variables are themselves negatively associated [9, 28], that is, the $\mathcal{R}_x$ variables are also negatively associated. Using the Chernoff bound for negatively associated random variables we have that

$$\Pr\left[\sum_{x\in A^+}\mathcal{R}_x \le 4\gamma d\right] \le \Pr\left[\sum_{x\in A^+}\mathcal{R}_x \le \frac{1}{2}\mathop{\mathbb{E}}_{S}\left[\sum_{x\in A^+}\mathcal{R}_x\right]\right] \le \exp(-12d\gamma) \le \frac{1}{2},$$

where in the penultimate transition we use Claim 4.8, and in the last transition we use the fact that $d > \frac{1}{\gamma}$. ∎

Next, condition on the the events of Claim 4.9, i.e., fix the set of samples S that were drawn from $\mathcal{D}'$ and assume that $\sum_{x\in A^+}\mathcal{R}_x \ge 4d\gamma$. Furthermore, fix the set

$$\Gamma_S \coloneqq \{x \in A^+ \mid \mathbb{R}_x = 1\}.$$

Consider a set of fresh samples $S' \sim (\mathcal{D}')^m$ and let $\hat{p}'_x$ denote the new empirical distribution. Let $Q'_x = 1$ if $p'_x > \frac{1}{d}$.

Claim 4.10. *Conditioned on the events of Claim 4.9, we have*

$$\Pr_{S\sim(\mathcal{D}')^m}\left[\sum_{x\in\Gamma_S} Q'_x > d\gamma\right] \ge \frac{1}{2}.$$

Proof of Claim. Note that since $\hat{p}'_x$ is a binomial variable with mean at least $\frac{1}{d}$, we have that $\Pr_{S'}[Q'_x] \ge \frac{1}{2}$ for any number of samples m [10].[8] Therefore, given Γ_S, we have

$$\mathop{\mathbb{E}}_{S'}\left[\sum_{x\in\Gamma_S} Q'_x\right] \ge \frac{1}{2}|\Gamma_S| \ge 2\gamma d.$$

Using the fact that Q'_x are negatively associated variables, we have

$$\Pr\left[\sum_{x\in\Gamma_S} Q'_x \le \gamma d\right] \le \exp\left(-6\gamma d\right) \le \frac{1}{2},$$

where the last transition is by the fact that $6d\gamma \ge 1$. Therefore, with constant probability over the choice of S', we have that $\sum_{x\in\Gamma_S} Q'_x > d\gamma$. ∎

Now consider the empirical cost-minimizers

$$f_{X_1} \in \operatorname*{argmin}_{f_X\in\mathcal{F}} \mathrm{cost}_S(f_X),$$

and

$$f_{X_2} \in \operatorname*{argmin}_{f_X\in\mathcal{F}} \mathrm{cost}_{S'}(f_X),$$

Notice that for any x with $\mathcal{R}_x = 1$, we have that $x \notin X_1$ and for any x with $Q'_x = 1$, we have that $x \in X_2$. Using Claims 4.9 and 4.10, with probability at least $1/4$ there are more than γd such x for which $\mathcal{R}_x = 1$ and $Q'_x = 1$, so we have that $\Delta_{\mathcal{D}}(f_{X_1}, f_{X_2}) > \gamma$ at probability at least $1/4$. ∎

[7] Similarly, for the case of $\mathcal{X}^-$, the corresponding variables would *all* be monotonically increasing.
[8] Here we are assuming for ease of exposition that m is not divisible by d, so $p'_x > \frac{1}{d} \Leftrightarrow p'_x \ge \frac{1}{d}$.

5 DISCUSSION

Our results show that polarization that arises from differences in subjective opinion is unlike polarization that arises from difficulty in processing objective information. Indeed, the Subjective Mixture Model is pessimistic in that polarization seems inevitable. By contrast, in the Objective Cost Model, our main result (Theorem 4.3) is positive: even though polarization arises, we can always introduce a slight bias into the information selection process (i.e., perturb the distribution) in a way that leads to consensus. While our model is admittedly a stylized abstraction of reality, the conceptual message remains appealing: in some situations, very mild interventions can be effective in preventing polarization.

The main shortcoming of Theorem 4.3, especially in terms of making this message more practical, is that the intervention needs to be tailored to the instance. It requires intimate knowledge of the hypothesis class, the cost function, and, for that matter, the composition of the set $\mathcal{F}_{\mathcal{D}}^{\epsilon}$. An important open question, therefore, is whether it is possible to design this intervention through a simple and tractable algorithm that does not explicitly construct $\mathcal{F}_{\epsilon}^{\mathcal{D}}$.

Let us end on a somewhat grandiose note. In recent years, economists have started to embrace machine learning. We view the interaction between the two fields as one of the greatest opportunities in the economics and computation area more broadly. In particular, we hope that our work will set the stage for a re-evaluation of social learning models (see Section 1.2) through the lens of machine learning.

REFERENCES

[1] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Ruth Urner. 2015. Efficient Learning of Linear Separators under Bounded Noise. In *Prceedings of the 28th Conference on Computational Learning Theory (COLT)*. 167–190.

[2] Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. 2016. Learning and 1-bit Compressed Sensing under Asymmetric Noise. In *Prceedings of the 29th Conference on Computational Learning Theory (COLT)*. 152–192.

[3] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. 2005. Theory of Classification: A Survey of Recent Advances. *ESAIM: Probability and Statistics* 9 (2005), 323–375.

[4] Aaron Fisher, C. Rudin, and Francesca Dominici. 2018. All Models are Wrong but Many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. arXiv:1801.01489. (2018).

[5] Roland G Fryer, Philipp Harms, and Matthew O. Jackson. 2019. Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *in press: Journal of the European Economic Association* (2019).

[6] Benjamin Golub and Evan Sadler. 2016. Learning in social networks. In *The Oxford Handbook of the Economics of Networks*, Yann Bramoullé, Andrea Galeotti, and Brian Rogers (Eds.). Oxford University Press, 504–542.

[7] Sanjeev Goyal. 2011. Learning in networks. In *Handbook of Social Economics*, Jess Benhabib, Alberto Bisin, and Matthew O. Jackson (Eds.). Vol. 1. Elsevier, 679–727.

[8] Matthew O. Jackson. 2019. *The Human Network: How Your Social Position Determines Your Power, Beliefs, and Behaviors*. Pantheon Books: New York.

[9] Kumar Joag-Dev and Frank Proschan. 1983. Negative association of random variables with applications. *The Annals of Statistics* (1983), 286–295.

[10] K. Jogdeo and S. Samuels. 1968. Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. *Annals of Mathematical Statistics* 39 (1968), 1191–1195.

[11] Thomas M. Liggett. 1985. *Interacting Particle Systems*. Springer.

[12] Charles G Lord, Lee Ross, and Mark R Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology* 37, 11 (1979), 2098.

[13] Enno Mammen and Alexandre B. Tsybakov. 1999. Smooth Discrimination Analysis. *The Annals of Statistics* 27, 6 (1999), 1808–1829.

[14] Charles McCoy. 2017. Anti-vaccination beliefs don't follow the usual political polarization. The Conversation, available from https://goo.gl/4gMCQP. (2017).

[15] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27 (2001), 415–444.

[16] George A Miller. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review* 101, 2 (1956), 343–352.

[17] Markus Mobius and Tanya Rosenblat. 2014. Social learning in economics. *Annual Review of Econonmics* 6, 1 (2014), 827–847.
[18] Sendhil Mullainathan and Jann Spiess. 2017. Machine learning: An applied econometric approach. *Journal of Economic Perspectives* 31, 2 (2017), 87–106.
[19] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna M. Wallach. 2018. Manipulating and Measuring Model Interpretability. arXiv:1802.07810. (2018).
[20] Ariel Rubinstein. 1998. *Modeling bounded rationality.* MIT Press.
[21] Grant Schoenebeck and Fang-Yi Yu. 2018. Consensus of Interacting Particle Systems on Erdös-Rényi Graphs. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA).* 1945–1964.
[22] Herbert A. Simon. 1947. *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization.* Macmillan.
[23] Christopher A Sims. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50, 3 (2003), 665–690.
[24] Lones Smith and Peter Sørensen. 2000. Pathological outcomes of observational learning. *Econometrica* 68, 2 (2000), 371–398.
[25] Daniel A Spielman and Shang-Hua Teng. 2004. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM* 51, 3 (2004), 385–463.
[26] Isaac Stanley-Becker. 2019. Officials in anti-vaccination 'hotspot' near Portland declare an emergency over measles outbreak. The Washington Post, available from https://goo.gl/EPP8Da. (2019).
[27] Federico Vazquez, Pavel L. Krapivsky, and Sidney Redner. 2003. Freezing and slow evolution in a constrained opinion dynamics model. *Journal of Physics A* 36 (2003), L61–L68.
[28] David Wajc. 2017. Negative Association — Definition, Properties, and Applications. Manuscript, available from https://goo.gl/j2ekqM. (2017).
[29] Andrea Wilson. 2014. Bounded memory and biases in information processing. *Econometrica* 82, 6 (2014), 2257–2294.
[30] Neal Young. 2012. Reverse Chernoff Bound. Theoretical Computer Science Stack Exchange, available from: https://cstheory.stackexchange.com/a/14476. (2012).

A MIND THE GAP

As shown in Theorem 4.6, the number of samples necessary to avoid polarization is at least $\Omega(\alpha^{-2}\ln(\gamma^{-1})\mathrm{VCD}(\mathcal{F}))$. This bound increases very slowly with the parameter γ that bounds the total amount of disagreement between two agents. By contrast, our upper bound of $O\left(\gamma^{-4}\alpha^{-2}\mathrm{VCD}(\mathcal{F})\right)$ in Theorem 4.3 shows a much higher dependence on $\frac{1}{\gamma}$. While we do not know which of these bounds is closer to 'reality,' it is important to note that the analysis of the lower bound on sample complexity is tight for the 'hard' example constructed in Theorem 4.6. So it is interesting to better understand where and how our upper bound analysis can be improved.

Looking back at the proof of Theorem 4.3, it consists of (1) the construction of $\widetilde{\mathcal{D}}$ with small $\mathrm{diam}(\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}})$ (in Lemma 4.4), and (2) sample complexity analysis of learning over hypotheses in $\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$ (in Lemma 4.5). We argue that the gap in the sample complexity of the 'hard' example is attributed to the latter part of the analysis.

To establish this, we show that Lemma 4.4 is tight for the construction of $\mathcal{F}$ and $\mathcal{D}$ in Theorem 4.6. That is, for any $\widetilde{\mathcal{D}}$ that is within distance α of $\mathcal{D}$ and has the same conditional label distributions as $\mathcal{D}$, we have that $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}) \geq \frac{2\epsilon}{\alpha}$. Indeed, note that for any $\widetilde{\mathcal{D}}$ with densities p_x for $x \in \mathcal{X}$, we have that

$$\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}} = \left\{ f_X \;\middle|\; \sum_{x \in \mathcal{X}^+ \setminus X} \left(p_x - \frac{1}{d}\right) + \sum_{x \in \mathcal{X}^- \cap X} \left(\frac{1}{d} - p_x\right) \leq \epsilon \right\}, \tag{18}$$

where, recall, we define

$$\mathcal{X}^+ := \{x \in \mathcal{X} \mid p_x \geq 1/d\}$$

and

$$\mathcal{X}^- := \{x \in \mathcal{X} \mid p_x < 1/d\}.$$

Note that by the definition of the L_1 distance between two distributions we have that

$$\sum_{x \in \mathcal{X}^+} \left(p_x - \frac{1}{d} \right) = \sum_{x \in \mathcal{X}^-} \left(\frac{1}{d} - p_x \right) = \frac{1}{2} \| \widetilde{\mathcal{D}} - \mathcal{D} \|_1 \leq \frac{\alpha}{2}.$$

Therefore, using Markov's inequality, there are sets $K^+ \subseteq \mathcal{X}^+$ and $K^- \subseteq \mathcal{X}^-$ of size $2\epsilon|\mathcal{X}^+|/\alpha$ and $2\epsilon|\mathcal{X}^-|/\alpha$, respectively, such that

$$\sum_{x \in K^+} \left(p_x - \frac{1}{d} \right) \leq \epsilon \quad \text{and} \quad \sum_{x \in K^-} \left(\frac{1}{d} - p_x \right) \leq \epsilon.$$

By definition of $\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$ in Equation (18), $f_{\mathcal{X}^+ \setminus K^+}$ and $f_{\mathcal{X}^+ \cup K^-}$ both belong to $\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$. Therefore,

$$\operatorname{diam}_{\mathcal{D}} \left(\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}} \right) \geq \Delta_{\mathcal{D}} \left(f_{\mathcal{X}^+ \setminus K^+}, f_{\mathcal{X}^+ \cup K^-} \right) = \frac{|K^-| + |K^+|}{d} \geq \frac{2\epsilon}{\alpha}.$$

As we see above, the gap between the upper and lower bound can be attributed to the sample complexity analysis of Lemma 4.5. Indeed, the analysis of Lemma 4.5 uses uniform convergence over the set of all hypotheses $\mathcal{F}$, i.e., $|\operatorname{err}_{\mathcal{D}}(f) - \operatorname{err}_S(f)| \leq \epsilon$ for all $f \in \mathcal{F}$. It is not hard to see that milder convergence properties are sufficient in this case. For example, if $\mathcal{F}_{\epsilon/2}^{\widetilde{\mathcal{D}}}$ is non-empty, then it is sufficient that there *exist* $f \in \mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$ such that

$$\operatorname{err}_S(f) \leq \operatorname{err}_{\mathcal{D}}(f) + \frac{\epsilon}{2}$$

as long as for all $f \in \mathcal{F} \setminus \mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$,

$$\operatorname{err}_S(f) > \operatorname{err}_{\mathcal{D}}(f) - \frac{\epsilon}{2}.$$

Additionally, it may be possible that $\mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$ and $\mathcal{F} \setminus \mathcal{F}_\epsilon^{\widetilde{\mathcal{D}}}$ have structural properties, yet unexplored, that yield better statistical learning properties. We leave the study of such properties and stronger upper (or possibly lower) bounds to future work.