A Unified Characterization of Private Learnability via Graph Theory

Noga Alon   NALON@MATH.PRINCETON.EDU
Department of Mathematics, Princeton University

Shay Moran   SMORAN@TECHNION.AC.IL
Departments of Mathematics, Computer Science, and Data and Decision Sciences, Technion
Google Research

Hilla Schefler   HILLAS@CAMPUS.TECHNION.AC.IL
Department of Mathematics, Technion

Amir Yehudayoff   AMIR.YEHUDAYOFF@GMAIL.COM
Department of Mathematics, Technion

Editors: Shipra Agrawal and Aaron Roth

Abstract

We provide a unified framework for characterizing pure and approximate differentially private (DP) learnability. The framework uses the language of graph theory: for a concept class $\mathcal{H}$, we define the contradiction graph G of $\mathcal{H}$. Its vertices are realizable datasets and two datasets S, S' are connected by an edge if they contradict each other (i.e., there is a point x that is labeled differently in S and S'). Our main finding is that the combinatorial structure of G is deeply related to learning $\mathcal{H}$ under DP. Learning $\mathcal{H}$ under pure DP is captured by the fractional clique number of G. Learning $\mathcal{H}$ under approximate DP is captured by the clique number of G. Consequently, we identify graph-theoretic dimensions that characterize DP learnability: the *clique dimension* and *fractional clique dimension*. Along the way, we reveal properties of the contradiction graph which may be of independent interest. We also suggest several open questions and directions for future research.

Keywords: Differential privacy, PAC learning, Contradiction graph, Clique number, Chromatic number, Fractional clique number, Fractional chromatic number, LP duality.

1. Introduction

Modern machine learning applications often involve handling sensitive data. Differential privacy (DP) (DMNS06) has emerged as a sound theoretical approach to reason about privacy in a precise and quantifiable fashion and has become the gold standard of statistical data privacy (DR14). It has also been implemented in practice, notably by Google (EPK14), Apple (app16b; app16a), and in the 2020 US census (DLS+). These developments raise the question:

Which learning tasks can be performed subject to differential privacy?

Extensive research has been carried out on this question within the framework of the classical Probably Approximately Correct (PAC) model (VC68; Val84), leading to the development of various characterizations of private learnability. Beimel, Nissim, and Stemmer introduced a quantity called the representation dimension that characterizes pure DP learnability (BNS13; BNS19). In a follow-up work, Feldman and Xiao found an interesting connection with communication complexity by associating every concept class $\mathcal{H}$ with a communication task whose complexity characterizes whether $\mathcal{H}$ is pure DP learnable (FX15).

Extensive research has also been devoted to studying the question of which learning tasks can be performed subject to approximate differential privacy, which is comparatively less demanding than pure differential privacy. Several characterizations of PAC learnability under this less restrictive version have been proven, including finite Littlestone dimension and online learnability (ALMM19; BLM20; ABL$^+$22), replicability and reproducibility (ILPS22; BGH$^+$23), low information complexity, PAC Bayes stability, and other variants of algorithmic stability (LM20; PNG22). For a more detailed discussion, please refer to (MM22).

Our Contribution. While the definitions of pure and approximate DP are closely related, the characterizations of learning under these distinct privacy constraints differ significantly. In this work, we devise a unified approach for characterizing both pure and approximate DP learnability. Our framework is based on graph theory; in particular, it demonstrates a tight link between private learnability and cliques in certain graphs, which we call *contradiction graphs*.

A clique in a graph G is a set δ of vertices such that every pair of distinct vertices in δ is connected by an edge. A fractional clique is a standard LP relaxation of a clique. A function $\delta : V \to [0, 1]$ is a fractional clique if for every independent set $I \subseteq V$,

$$\sum_{v \in I} \delta(v) \leq 1.$$

The size of a fractional clique δ is the sum $\sum_{v \in V} \delta(v)$. Notice that if $\delta(v) \in \{0, 1\}$ for all v then δ is the indicator function of a clique, and its size is the number of vertices in the clique. The clique number of G, denoted $\omega(G)$, is the largest size of a clique in G. Similarly, the fractional clique number of G, denoted $\omega^\star(G)$, is the largest size of a fractional clique in G. Notice that $\omega^\star(G) \geq \omega(G)$.

Definition *[Contradiction Graph] Let $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ be a concept class and let $m \in \mathbb{N}$. The contradiction graph of order m of $\mathcal{H}$ is an undirected graph $G_m = G_m(\mathcal{H})$ whose vertices are datasets of size m that are consistent with $\mathcal{H}$. Two datasets are connected by an edge whenever they contradict each other.*

In other words, the vertices of the contradiction graph $G_m(\mathcal{H})$ are $\mathcal{H}$-realizable sequences of length m, and $\{S', S''\}$ is an edge if there is $x \in \mathcal{X}$ such that $(x, 0) \in S'$ and $(x, 1) \in S''$. Let $\omega_m = \omega(G_m)$ and $\omega_m^\star = \omega^\star(G_m)$ denote the clique and fractional clique numbers of G_m. We prove that both ω_m and $\omega_m^\star$ satisfy a polynomial-exponential dichotomy:

(i) For every $\mathcal{H}$, either $\omega_m^\star = 2^m$ for all m, or $\omega_m^\star \leq \texttt{poly}(m)$.
(ii) For every $\mathcal{H}$, either $\omega_m = 2^m$ for all m, or $\omega_m \leq \texttt{poly}(m)$.

These dichotomies characterize pure and approximate DP learnability:

(i) $\mathcal{H}$ is pure DP learnable if and only if $\omega_m^\star \leq \texttt{poly}(m)$.
(ii) $\mathcal{H}$ is approximately DP learnable if and only if $\omega_m \leq \texttt{poly}(m)$.

These characterizations yield graph-theoretic dimensions of $\mathcal{H}$ that characterize private learning. Define the clique dimension of $\mathcal{H}$, denoted by $\texttt{CD}(\mathcal{H})$, as the largest $m \in \mathbb{N} \cup \{\infty\}$ for which $\omega_m = 2^m$. Analogously, define the fractional clique dimension of $\mathcal{H}$, denoted by $\texttt{CD}^\star(\mathcal{H})$, as the largest m for which $\omega_m^\star = 2^m$. Thus,

(i) $\texttt{CD}^\star(\mathcal{H}) < \infty$ if and only if $\mathcal{H}$ is pure DP learnable.
(ii) $\texttt{CD}(\mathcal{H}) < \infty$ if and only if $\mathcal{H}$ is approximately DP learnable.

Technically, our proofs rely on the fact that the fractional clique and chromatic numbers are equal. For finite graphs, this fact follows from LP duality, however, in our setting the contradiction graph can be infinite. Lastly, we prove that the contradiction graph exhibits strong duality: the fractional clique and chromatic numbers are equal for every (possibly infinite) contradiction graph. This part is based on tools from functional analysis and topology.

Note that in this work we focus on providing a unified and concise framework to characterize and study private learning. Important topics, such as regarding the informational or computational complexity are not addressed in this work. Nevertheless, given that graph theory is a well-studied area with sophisticated tools and techniques, we hope that the equivalence established here will facilitate a deeper integration between learning theory and graph theory. Specifically, we anticipate that it will provide new tools and insights to tackle other fundamental questions in learning theory.

Organization. In Section 2 we present the main results in greater detail. Section 3 contains background and relevant definitions in learning theory, graph theory, and differential privacy, together with basic facts regarding the contradiction graph. In Section 4 we provide a technical overview of the proofs of the main theorems. Finally, in Section 5, we suggest directions for future work. Appendices A to D contain the full proofs.

2. Main Results

We use standard definitions and terminology from graph theory, learning theory, and differential privacy; see Section 3 for detailed definitions.

2.1. Dichotomies and Dimensions

In this section we present Theorems 2 and 3 which concern cliques and fractional cliques in the contradiction graph. These results are key in our characterizations of private learnability, but they might also be of independent interest as combinatorial results.

We begin with a basic lemma which shows that the contradiction graph does not contain cliques or fractional cliques of size larger than 2^m.

Lemma 1 *Let $\mathcal{H}$ be a class and $m \in \mathbb{N}$ and let ω_m and $\omega_m^\star$ denote the clique and fractional clique numbers of $G_m(\mathcal{H})$. Then, $\omega_m \leq \omega_m^\star \leq 2^m$.*

A short proof of this lemma is provided in Appendix D.

Theorem 2 (Clique Number) *Let $\mathcal{H}$ be a class and let ω_m denote the clique number of $G_m(\mathcal{H})$. Then, exactly one of the following statements holds:*
  1. $\omega_m = 2^m$ for all m.
  2. $\omega_m \leq P(m)$ for all m, where $P(m)$ is a polynomial.

Theorem 3 (Fractional Clique Number) *Let $\mathcal{H}$ be a class and let $\omega_m^\star$ denote the fractional clique number of $G_m(\mathcal{H})$. Then, exactly one of the following statements holds:*
  1. $\omega_m^\star = 2^m$ for all m.
  2. $\omega_m^\star \leq P(m)$ for all m, where $P(m)$ is a polynomial.

Theorems 2 and 3 motivate the following definitions.

Definition 4 (Clique Dimension) *The clique dimension of a concept class $\mathcal{H}$, denoted $\mathtt{CD}(\mathcal{H})$, is defined as follows:*

$$\mathtt{CD}(\mathcal{H}) := \sup\{m : \omega_m = 2^m\} \in \mathbb{N} \cup \{\infty\},$$

where ω_m is the clique number of the contradiction graph $G_m(\mathcal{H})$.

Definition 5 (Fractional Clique Dimension) *The fractional clique dimension of a concept class $\mathcal{H}$, denoted $\mathtt{CD}^\star(\mathcal{H})$, is defined as follows:*

$$\mathtt{CD}^\star(\mathcal{H}) := \sup\{m : \omega_m^\star = 2^m\} \in \mathbb{N} \cup \{\infty\},$$

where $\omega_m^\star$ is the fractional clique number of the contradiction graph $G_m(\mathcal{H})$.

Thus, Theorems 2 and 3 demonstrate that the clique and fractional clique dimensions satisfy a dichotomy similar to the Sauer-Shelah-Perles (SSP) dichotomy of the VC dimension (Sau72). Theorems 2 and 3 are key in our characterizations of private PAC learnability. This is analogous to the crucial role played by the SSP lemma in the characterization of PAC learnability.

It is worthwhile to note that the polynomial/exponential dichotomy in Theorem 3 is weaker than the one in Theorem 2. Specifically, in Theorem 2 the degree of the polynomial $P(m)$ is the clique dimension $\mathtt{CD}(\mathcal{H})$ (see Lemma 18). In contrast, our proof of Theorem 3 does not imply a bound on the degree of $P(m)$ in terms of the fractional clique dimension. Rather, the implied bound depends on the difference $2^m - \omega_m^\star > 0$, where m is any integer for which this difference is positive (m exists by Lemma 1). We leave as an open question to determine whether the fractional clique number $\omega_m^\star$ is upper bounded by a polynomial $P(m)$ whose degree depends <u>only</u> on the fractional clique dimension.

2.2. Private Learnability: Characterizations

We next present the characterizations of pure and approximate private learnability.

Theorem 6 (Pure DP Learnability) *The following statements are equivalent for a concept class $\mathcal{H}$.*
  1. *$\mathcal{H}$ is pure differentially private PAC learnable.*
  2. *$\mathcal{H}$ has finite fractional clique dimension.*

Theorem 7 (Approximate DP Learnability) *The following statements are equivalent for a concept class $\mathcal{H}$.*
  1. *$\mathcal{H}$ is approximately differentially private PAC learnable.*
  2. *$\mathcal{H}$ has finite clique dimension.*

Theorems 6 and 7 provide a unified characterization of private PAC learnability in terms of cliques and fractional cliques. Roughly speaking, these theorems assert that large (fractional) cliques in the contradiction graph correspond to tasks that are hard to learn privately. However, our proofs of Theorems 6 and 7 do not explicitly illustrate this correspondence. Instead, our proofs follow an indirect path by linking the clique and fractional clique dimensions to the representation and Littlestone dimensions, respectively. This implies the stated equivalences since the representation and Littlestone dimensions characterize pure and approximate DP learnability (BNS19; ABL$^+$22).

It would be interesting to find direct proofs that illustrate the correspondence between large cliques in the contradiction graphs and hard learning tasks. A clique of size k in the contradiction

graph is simply a set $\mathcal{S}$ of k realizable datasets such that every two distinct datasets $S', S'' \in \mathcal{S}$ disagree on some example. Analogously, a fractional clique of size k is a distribution ν over realizable data sets such that $\Pr_{S \sim \nu}[h \text{ is consistent with } S] \leq 1/k$ for every hypothesis h. Hence, it is quite natural to speculate that there exists a natural and direct conversion between cliques and fractional cliques, and realizable distributions on datasets that are hard for private learning.

Duality, Representation Dimension, and Communication Complexity. The fractional clique number is defined via a linear program and as such it has a dual definition. The latter is called the fractional chromatic number and is defined as follows. Let G be an undirected graph and let $\mathcal{I}$ denote the family of independent sets in G. A fractional coloring is an assignment $c : \mathcal{I} \to \mathbb{R}_{\geq 0}$ such that for every vertex v we have $\sum_{I:v \in I} c(I) \geq 1$. The number of colors in c is defined by $\mathtt{col}(c) = \sum_{I \in \mathcal{I}} c(I)$. The fractional chromatic number, denoted by $\chi^\star(G)$, is the infimum number of colors in a fractional coloring. Notice that $\chi^\star(G) \leq \chi(G)$, where $\chi(G)$ is the chromatic number of G. This holds because any coloring defines a fractional coloring by assigning 1 to each color class.

Fractional colorings of the contradiction graph have a natural learning theoretic interpretation. A fractional coloring c of $G_m(\mathcal{H})$ corresponds to a distribution μ over hypotheses such that for every realizable dataset S of size m:

$$\Pr_{h \sim \mu}[h \text{ is consistent with } S] \geq \frac{1}{\mathtt{col}(c)}. \tag{1}$$

The fractional chromatic number of the contradiction graph is tightly linked to the representation dimension (see Definition 9). The latter is a dimension introduced by (BNS19) to characterize pure DP learnability. Roughly, the representation dimension of $\mathcal{H}$ is the minimal integer d for which there exists a distribution $\mathcal{P}$ over hypothesis <u>classes</u> of size $\leq 2^d$ such that for every realizable distribution $\mathcal{D}$, a random class $H \sim \mathcal{P}$ contains with probability at least $3/4$ an hypothesis $h = h_\mathcal{D}$ whose loss with respect to $\mathcal{D}$ is at most $1/4$. The representation dimension can be interpreted as a lossy variant of the fractional chromatic number. Indeed, consider a uniformly sampled hypothesis h drawn from a random class $H \sim \mathcal{P}$, and pick $\mathcal{D}$ to be the uniform distribution over the examples in a realizable dataset S. Thus, with probability at least $\frac{3}{4}2^{-d}$, the random hypothesis h classifies correctly $3/4$ of the examples in S.

A similar link exists between the fractional clique number of $G_m(\mathcal{H})$ and the communication complexity theoretic characterization of pure private learnability by (FX15). The latter is based on a communication game between two players whom we call Alice and Bob. In the game, Alice's input is a hypothesis $h \in \mathcal{H}$, Bob's input is a point x, and Alice sends a <u>single</u> message to Bob. Their goal is that with probability at least $3/4$, Bob will be able to decode $h(\overline{x})$ from Alice's message. Feldman and Xiao showed that the optimal number of bits required to perform this task is proportional to the representation dimension of $\mathcal{H}$, and hence characterizes pure DP learnability.

Fractional cliques are linked to hard distributions for the above communication game. A fractional clique δ corresponds to a distribution ν over realizable datasets S of size m such that for every hypothesis h:

$$\Pr_{S \sim \nu}\Big[h \text{ is consistent with } S\Big] \leq \frac{1}{|\delta|}. \tag{2}$$

Consider a variant of the above equation where the event "h is consistent with S" is replaced by the event "h classifies correctly at least $3/4$ of the examples in S". This variant induces a hard distribution for the communication game as follows. Pick a random dataset $S \sim \nu$ and a concept

$h \in \mathcal{H}$ such that h is consistent with S. Let Alice's input be h and Bob's input be a random (unlabeled) example from S. Standard arguments in communication complexity show that this distribution is indeed hard for the communication game.

To summarize the discussion, LP duality implies that the fractional clique and chromatic numbers are equal. When considering this along with the aforementioned connections, it illuminates a dual relationship between the representation dimension and the communication-complexity-based characterization of pure private learnability.

2.3. Strong Duality in Infinite Contradiction Graphs

Our proofs heavily rely on the equality of the fractional clique and chromatic numbers. For finite graphs, this equality is a consequence of LP duality. However, in learning theory, we often study infinite hypothesis classes $\mathcal{H}$ whose contradiction graphs are therefore also infinite. In general, LP duality does not apply in infinite dimensional spaces. Therefore, we prove the next theorem showing that any (possibly infinite) contradiction graph $G_m(\mathcal{H})$ satisfies that its fractional clique and chromatic numbers are equal and are bounded from above by 2^m.

Theorem 8 *Let $\mathcal{X}$ be an arbitrary domain, $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ a concept class and $m \in \mathbb{N}$. Let $\omega_m^{\star}$ and $\chi_m^{\star}$ denote the fractional clique and chromatic numbers of the contradiction graph $G_m(\mathcal{H})$. Then,*

$$\omega_m^{\star} = \chi_m^{\star} \leq 2^m.$$

Moreover, there exists a fractional coloring realizing $\chi_m^{\star}$. (I.e. the infimum is in fact a minimum.)

The proof of Theorem 8 uses tools from functional analysis, topology, and measure theory. The equality $\omega_m^{\star} = \chi_m^{\star}$ is derived using Sion's Theorem (Sio58), and the upper bound of 2^m hinges on Kolmogorov's Extension Theorem (see, e.g. Theorem 2.4.3 in (Tao11)). This proof appears in Appendix C, which we attempted to present in a manner accessible for readers who may only have basic familiarity with topology and functional analysis.

3. Preliminaries

3.1. Learning

PAC Learning. We use standard notations from statistical learning; for more details see e.g. (SSBD14). Let $\mathcal{X}$ be a domain; for simplicity, in this work, we assume that $\mathcal{X}$ is countable, although our arguments apply more generally. Given an hypothesis $h : \mathcal{X} \to \{0,1\}$, the *empirical loss* of h with respect to a dataset $S = \big((x_1, y_1), \ldots, (x_m, y_m)\big)$ is defined as $\mathsf{L}_S(h) \coloneqq \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[h(x_i) \neq y_i]$. We say that h is *consistent* with S if $\mathsf{L}_S(h) = 0$. The *population loss* of h with respect to a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ is defined as $\mathsf{L}_{\mathcal{D}}(h) \coloneqq \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$. A distribution $\mathcal{D}$ over labeled examples is *realizable* with respect to $\mathcal{H}$ if $\inf_{h\in\mathcal{H}} \mathsf{L}_{\mathcal{D}}(h) = 0$

For a set Z, let $Z^{\star} = \cup_{n=0}^{\infty} Z^n$. A *learning rule* $\mathcal{A}$ is a (possibly randomized) algorithm that takes as input a dataset $S \in (\mathcal{X} \times \{0,1\})^{\star}$ and outputs a hypothesis $h = \mathcal{A}(S) \in \{0,1\}^{\mathcal{X}}$. In the PAC learning model, the input S is sampled i.i.d. from a realizable distribution $\mathcal{D}$, and the learner's goal is to output a hypothesis with small population loss with respect to $\mathcal{D}$. More precisely, let $m, \alpha, \beta > 0$. We say that an algorithm $\mathcal{A}$ is an (m, α, β)-learner for $\mathcal{H}$ if for every realizable distribution $\mathcal{D}$, $\Pr_{S\sim\mathcal{D}^m}[\mathsf{L}_{\mathcal{D}}(\mathcal{A}(S)) > \alpha] < \beta$. Here, α is called the *error*, β is the *confidence*

parameter, and m is the *sample complexity*. A class $\mathcal{H}$ is *PAC learnable* if there exists vanishing $\alpha(m), \beta(m) \to 0$ and an algorithm $\mathcal{A}$ such that for all m, algorithm $\mathcal{A}$ is a $(m, \alpha(m), \beta(m))$-learner for $\mathcal{H}$.

Differential Privacy. We use standard notations from differential privacy literature; for more details see e.g. (DR14; Vad17). A randomized learning algorithm $\mathcal{A}$ is *differentially private* with parameters (ϵ, δ), if for every input datasets $S, S' \in (\mathcal{X} \times \{0, 1\})^m$ that differ on a single example, and every event $E \subseteq \{0, 1\}^{\mathcal{X}}$:

$$\Pr[\mathcal{A}(S) \in E] \leq e^\epsilon \Pr[\mathcal{A}(S') \in E] + \delta,$$

where the probability is over the randomness of $\mathcal{A}$. The parameters ϵ and δ are usually treated as follows: ϵ is a small constant (say ≤ 0.1) and δ is negligible, $\delta = m^{-\omega(1)}$ where m is the input datasets size. When $\delta = 0$ we say that $\mathcal{A}$ is *pure differentially private*, and when $\delta > 0$ we say that $\mathcal{A}$ is *approximate differentially private*.

An hypothesis class $\mathcal{H}$ is *pure privately learnable* (abbreviated *pure DP learnable*) if it is privately learnable by an algorithm which is $(\epsilon(m), 0)$-differentially private, where $\epsilon(m) = O(1)$ is a numerical constant. An hypothesis class $\mathcal{H}$ is *approximately privately learnable* (abbreviated *DP learnable*) if it is PAC learnable by an algorithm $\mathcal{A}$ which is $(\epsilon(m), \delta(m))$-differentially private, where $\epsilon(m) = O(1)$ is a numerical constant and $\delta(m) = m^{-\omega(1)}$.

Representation Dimension. The representation dimension is a combinatorial parameter introduced by Beimel et al. (BNS19) who used it to characterize pure DP learnability.

Definition 9 (Representation Dimension (BNS19)) *The* representation dimension *of a concept class* $\mathcal{H}$, *denoted* $\mathtt{RepDim}(\mathcal{H})$, *is defined to be* $\ln(d)$, *where* d *is the minimal integer for which there exists a distribution* $\mathcal{P}$ *over hypothesis classes of size* d *that satisfies the following. For every distribution* $\mathcal{D}$ *on labeled examples that is realizable by* $\mathcal{H}$,

$$\Pr_{\mathcal{C} \sim \mathcal{P}} \left[\exists h \in \mathcal{C} \text{ s.t. } L_{\mathcal{D}}(h) \leq \frac{1}{4} \right] \geq \frac{3}{4}.$$

As (BNS19) show, the constants $1/4, 3/4$ above can be replaced by any other pair of constants in $(0, 1)$ without changing the semantics of the definition. This follows from the next lemma:

Lemma 10 (Boosting Probabilistic Representation, Lemma 18 (BNS19)) *Let* $\mathcal{H}$ *be a class with* $\mathtt{RepDim}(\mathcal{H}) = d < \infty$. *Then for every* $0 < \alpha, \beta < 1$ *there exists a probability distribution* $\mathcal{P}$ *over hypothesis classes of size* $O\left((\frac{1}{\alpha})^{d + \ln\ln\ln(\frac{1}{\alpha}) + \ln\ln(\frac{1}{\beta})} \right)$ *which satisfies the following. For every realizable distribution* $\mathcal{D}$ *on labeled examples which is realizable by* $\mathcal{H}$,

$$\Pr_{\mathcal{C} \sim \mathcal{P}} [\exists h \in \mathcal{C} \text{ s.t. } L_{\mathcal{D}}(h) \leq \alpha] \geq 1 - \beta.$$

Littlestone Dimension. The Littlestone dimension is a combinatorial parameter that captures mistake and regret bounds in online learning (Lit88; BDPSS09). A *mistake tree* is a binary decision tree whose nodes are labeled with instances from $\mathcal{X}$ and whose edges are labeled by 0 or 1 such that each internal node has one outgoing edge labeled 0 and one outgoing edge labeled 1. A root-to-leaf path in a mistake tree is a sequence of labeled examples $(x_1, y_1), \ldots, (x_d, y_d)$. The point x_i is the label of the i'th internal node in the path, and y_i is the label of its outgoing edge to the next node in the path. A class $\mathcal{H}$ *shatters* a mistake tree if every root-to-leaf path is realizable by $\mathcal{H}$. The *Littlestone dimension* of $\mathcal{H}$, denoted $\mathtt{LD}(\mathcal{H})$, is the largest number $d \in \mathbb{N} \cap \{\infty\}$ such that there exists a

complete binary mistake tree of depth d shattered by $\mathcal{H}$. If $\mathcal{H}$ shatters arbitrarily deep mistake trees then we write $\texttt{LD}(\mathcal{H}) = \infty$.

3.2. Graph Theory

Cliques, Colorings, and Distributions. Let $G = (V, E)$ be a (possibly countable) graph. Denote by $\omega(G)$ the *clique number* of G, which is largest size of a clique in G. Denote by $\chi(G)$ the *chromatic number* of G, which is the smallest number of colors needed to color the vertices of G so that no two adjacent vertices share the same color. The clique and chromatic numbers have natural LP relaxations.

A *fractional clique* is a function $\delta : V \to \mathbb{R}_{\geq 0}$ such that $\sum_{v \in I} \delta(v) \leq 1$ for every independent set I. The *size* of δ is $|\delta| := \sum_{v \in V} \delta(v)$. The *fractional clique number* of G, denoted $\omega^\star(G)$, is defined by $\omega^\star(G) := \sup_\delta |\delta|$. A *fractional coloring* is a finite measure[1] c on $\mathcal{I}$, where $\mathcal{I}$ is the family of all independent sets in G, such that $c(\{I : v \in I\}) \geq 1$ for every $v \in V$. The *fractional chromatic number* of G, denoted $\chi^\star(G)$, is defined by $\chi^\star(G) := \inf_c c(\mathcal{I})$. Note that from (weak) LP-duality, for any graph G, $\omega(G) \leq \omega^\star(G) \leq \chi^\star(G) \leq \chi(G)$. In finite graphs strong LP-duality holds and $\omega^\star(G) = \chi^\star(G)$. Theorem 8 extends this equality to any (possibly infinite) contradiction graph.

There is a natural correspondence between fractional colorings of a graph G and distributions over independent sets. Given a fractional coloring c, normalizing c by the number of colores in c, denoted by $\texttt{col}(c) := c(\mathcal{I})$, induces a distribution μ over independent sets, such that for every vertex v,

$$\Pr_{I \sim \mu}[v \in I] = \frac{1}{\texttt{col}(c)} \cdot c(\{I : v \in I\}).$$

Define the *value* of a distribution μ over $\mathcal{I}$ to be $\texttt{val}(\mu) := \inf_{v \in V} \Pr_{I \sim \mu}[v \in I]$. By taking infimum,

$$\texttt{val}(\mu) := \inf_{v \in V} \Pr_{I \sim \mu}[v \in I] = \frac{1}{\texttt{col}(c)}.$$

From the other direction, given a distribution μ over $\mathcal{I}$, normalizing μ by $\texttt{val}(\mu)$ induces a fractional coloring of G. Minimizing the number of colors of a fractional coloring is equivalent to maximizing the value of the corresponding distribution, hence

$$\frac{1}{\chi^\star(G)} = \sup_\mu \texttt{val}(\mu) = \sup_\mu \inf_{v \in V} \Pr_{I \sim \mu}[v \in I].$$

Similarly, there is a correspondence between fractional cliques of a graph G and distributions over vertices. Normalizing a fractional clique $\delta : V \to \mathbb{R}_{\geq 0}$ by $|\delta|$, induces a distribution ν over vertices, and normalizing a distribution ν over V by $\texttt{val}(\nu) := \sup_{I \in \mathcal{I}} \Pr_{v \sim \nu}[v \in I]$ induces a fractional clique. Similarly,

$$\frac{1}{\omega^\star(G)} = \inf_\nu \texttt{val}(\nu) = \inf_\nu \sup_{I \in \mathcal{I}} \Pr_{v \sim \nu}[v \in I].$$

For further reading about fractional graph theory see (SU13).

---

1. For a finite G, one typically defines a fractional coloring as a function $c : \mathcal{I} \to \mathbb{R}_{\geq 0}$ such that $\sum_{v \in I} c(I) \geq 1$ for every $v \in V$. The latter amounts to c being a discrete measure on $\mathcal{I}$. In this work, we also consider infinite graphs and use this more general definition which allows for non-discrete measures on $\mathcal{I}$. We refer the reader to Section C for a more detailed discussion.

3.3. The Contradiction Graph: Basic Facts

In this section, we state basic lemmas about the structure of the contradiction graph. We begin with a discussion about the relation between interpolating learning rules and colorings of the contradiction graph. Then, we discuss the relation between fractional colorings and cliques, and distributions over hypotheses and realizable datasets. Omitted proofs can be found at Appendix D.

Interpolating Algorithms are Proper Colorings. There is a correspondence between proper colorings of the contradiction graph and interpolating learning rules. A deterministic learning rule $\mathcal{A}$ is said to be *interpolating* with respect to a class $\mathcal{H}$ if for every realizable input dataset S, the output hypothesis $h := \mathcal{A}(S)$ satisfies $h(x) = y$ for every labeled example $(x, y) \in S$. A *proper coloring* of a graph is an assignment of a color to each vertex so that no two adjacent vertices share the same color. In other words, a coloring is a partition of the vertices such that every subset in the partition is an independent set. The correspondence between colorings and interpolation algorithms is a direct result of the following lemma which identifies independent sets in the contradiction graph with hypotheses[2].

Lemma 11 (Independent sets and consistent hypotheses) *Let $\mathcal{H}$ be a class and m be a natural number.*

1. *For every independent set I in $G_m(\mathcal{H})$, there exists an hypothesis $h \in \{0, 1\}^{\mathcal{X}}$ such that h is consistent with every dataset $S \in I$; i.e. for every dataset $S \in I$ and every example $(x, y) \in S$, we have $h(x) = y$.*
2. *For every hypothesis h, the set of all datasets of size m that are consistent with h is an independent set in $G_m(\mathcal{H})$; i.e. the set*

$$V_h := \{S = ((x_1, y_1), \ldots, (x_m, y_m)) \in V_m(\mathcal{H}) \mid \forall i, h(x_i) = y_i\}$$

is independent in $G_m(\mathcal{H})$.

Colorings $\rightarrow$ Algorithms. Let S be a realizable dataset of size m, and consider a coloring of $G_m(\mathcal{H})$. By part 1 of Lemma 11 there exists a hypothesis h which is consistent with every dataset in $G_m(\mathcal{H})$ colored with the same color as S. Now simply define $\mathcal{A}(S) = h$. This defines an interpolating learning rule $\mathcal{A}$ for $\mathcal{H}$.

Algorithms $\rightarrow$ Colorings. Given an interpolating learning rule $\mathcal{A}$, by part 2 of Lemma 11 for every realizable dataset S of size m, the set $V_{\mathcal{A}(S)}$ is independent. Note that since $\mathcal{A}$ is interpolating every dataset $S \in V_m(\mathcal{H})$ is covered, indeed $S \in V_{\mathcal{A}(S)}$. Next, assign a unique color to all datasets in $V_{\mathcal{A}(S)}$. (If there is more than one possible color option for some datasets then arbitrarily choose one.) This defines a proper coloring of $G_m(\mathcal{H})$.

Fractional Cliques and Colorings of the Contradiction Graph. In the contradiction graph independent sets correspond to hypothesis and vertices are realizable datasets. Therefore, fractional colorings can be viewed as distributions over hypotheses, and fractional cliques as distributions over realizable datasets:

Lemma 12 (Fractional cliques and colorings vs. distributions) *Let $\mathcal{H}$ be a class, $m \in \mathbb{N}$. Then,*

---

2. The mapping described in Lemma 11 is $1 - 1$ for maximal independent sets.

1. *There exists a fractional coloring c of $G_m(\mathcal{H})$ with $\texttt{col}(c) = \alpha > 0$ if and only if there exists a distribution μ over hypotheses such that*

$$\inf_S \Pr_{h \sim \mu} [h \text{ is consistent with } S] = \frac{1}{\alpha},$$

where the infimum is taken over realizable datasets of size m.

2. *There exists a fractional clique δ of $G_m(\mathcal{H})$ with $|\delta| = \alpha > 0$ if and only if there exists a distribution ν over realizable datasets of size m such that*

$$\sup_h \Pr_{S \sim \nu} [h \text{ is consistent with } S] = \frac{1}{\alpha},$$

where the supremum is taken over hypotheses $h \in \{0,1\}^{\mathcal{X}}$.

Let $\omega_m^\star$ and $\chi_m^\star$ denote the fractional clique and chromatic numbers of $G_m(\mathcal{H})$. It follows that

$$\frac{1}{\chi_m^\star} = \sup_\mu \inf_\nu \mathop{\mathbb{E}}_{\substack{h \sim \mu \\ S \sim \nu}} \left[\mathbb{1}[h \text{ is consistent with } S] \right], \tag{3}$$

$$\frac{1}{\omega_m^\star} = \inf_\nu \sup_\mu \mathop{\mathbb{E}}_{\substack{h \sim \mu \\ S \sim \nu}} \left[\mathbb{1}[h \text{ is consistent with } S] \right]. \tag{4}$$

where the supremum is taken over distributions over hypotheses, and the infimum is taken over distributions over realizable datasets of size m. (See discussion in Section 3.2.)

Corollary 13 *Let $\omega_m^\star$ denote the fractional clique number of $G_m(\mathcal{H})$. Then there exists a distribution $\mu^\star$ over hypotheses such that for every realizable dataset S of size m,*

$$\Pr_{h \sim \mu^\star} [h \text{ is consistent with } S] \geq \frac{1}{\omega_m^\star}.$$

Proof By Theorem 8 there exists a fractional coloring of value $\chi_m^\star = \omega_m^\star$, thus by Lemma 12 there exists a distribution $\mu^\star$ over hypotheses as wanted. ∎

4. Technical Overview

4.1. Dichotomies and Dimensions: Theorems 2 and 3

We begin with overviewing the proof of Theorem 2. Assuming there exists a natural number d such that $\omega_d < 2^d$, we need to show that $|\delta| \leq \texttt{poly}(m)$ for every m and for every clique δ in G_m. The crux of the proof is to show that there exists a *balanced* instance x_1 in the following sense:

$$|\{S \in \delta : (x_1, 0) \in S\}| \geq \Omega\left(\frac{|\delta|}{m}\right) \text{ and } |\{S \in \delta : (x_1, 1) \in S\}| \geq \Omega\left(\frac{|\delta|}{m}\right) \tag{5}$$

To see how Equation (5) completes the proof, let $\mathcal{H}_{x \to y} = \{h \in \mathcal{H} : h(x) = y\}$. Notice that at least one of $G_{d-1}(\mathcal{H}_{x_1 \to 0})$, $G_{d-1}(\mathcal{H}_{x_1 \to 1})$ does not contain a clique of size 2^{d-1}. Indeed, if δ_0, δ_1 are cliques of size 2^{d-1} in $G_{d-1}(\mathcal{H}_{x_1 \to 0})$ and $G_{d-1}(\mathcal{H}_{x_1 \to 1})$, then they can be combined to form a clique of size 2^d in $G_d(\mathcal{H})$ by adding the example $(x_1, 0)$ to every dataset in δ_0 and the example $(x_1, 1)$ to every dataset in δ_1, and taking their union.

Repeating this argument $d - 1$ times, we obtain a class $\mathcal{H}' = \mathcal{H}_{x_1 \to y_1, \dots x_{d-1} \to y_{d-1}}$ such that (i) $G_1(\mathcal{H}')$ does not contain a clique of size 2, and (ii) there is a clique in $G_m(\mathcal{H}')$ whose size is

at least $|\delta|/(c \cdot m)^{d-1}$ for some constant c. On the other hand, the first item means that $|\mathcal{H}'| = 1$ and hence every clique in $G_m(\mathcal{H}')$ has size 1. Thus, $1 \geq |\delta|/(c \cdot m)^{d-1}$, which implies that $|\delta| \leq (c \cdot m)^{d-1} = \texttt{poly}(m)$ as required.

We prove that there exists a balanced instance (i.e. that satisfies Equation (5)) constructively using a greedy procedure (see Lemma 17 for the short proof).

The proof of Theorem 3 is more involved and it integrates different ideas and techniques such as LP duality, probabilistic arguments, and regret analysis from online learning. Our objective is to demonstrate that if there exists a natural number d such that $\omega_d^\star < 2^d$, then $|\delta| \leq \texttt{poly}(m)$ for *every* fractional clique δ in G_m. The first step in the proof is to apply LP duality, which reduces the latter to show that *there exists* a fractional coloring of $G_m(\mathcal{H})$ that employs $\texttt{poly}(m)$ colors. Fractional colorings correspond to distributions over hypotheses, therefore, it suffices to prove the following statement: there exists a distribution μ over hypotheses such that for every dataset S of size m that is realizable by $\mathcal{H}$:

$$\Pr_{h \sim \mu}[h \text{ is consistent with } S] \geq 1/\texttt{poly}(m). \tag{6}$$

We obtain the distribution μ as follows. By the above correspondence, there is a distribution μ' such that $\Pr_{h \sim \mu'}[h \text{ is consistent with } S] \geq 2^{-d} + \epsilon$, for every realizable dataset S of size d. The distribution μ is obtained by independently sampling $\Theta(\log m/\epsilon^2)$ hypotheses from μ' and taking their majority vote. Interestingly, the analysis showing that μ satisfies Equation (6) follows by a reduction to online prediction using expert advice.

4.2. Characterizations of Private Learnability: Theorems 6 and 7

Our proof of Theorems 6 and 7 rely on the exponential-polynomial dichotomies (Theorems 3, and 2). We begin with overviewing the proof of Theorem 7. By (ABL$^+$22), a class $\mathcal{H}$ is approximately DP learnable if and only if it has a finite Littlestone dimension. Thus, it suffices to show that the clique dimension is finite if and only if the Littlestone dimension is finite. Our proof yields explicit bounds of

$$\texttt{LD}(\mathcal{H}) \leq \texttt{CD}(\mathcal{H}) \leq O(\texttt{LD}(\mathcal{H}) \log \texttt{LD}(\mathcal{H})),$$

where $\texttt{LD}(\mathcal{H})$ denotes the Littlestone dimension of $\mathcal{H}$. One direction is straightforward. If the Littlestone dimension is at least m, then there is a mistake tree of depth m that is shattered by $\mathcal{H}$. Each branch of the tree corresponds to a realizable dataset of length m, and the collection of 2^m datasets corresponding to the branches form a clique. Indeed, every two datasets disagree on the example corresponding to the least common ancestor of their branches.

The converse direction is more challenging. Assume that the clique dimension $\texttt{CD}(\mathcal{H})$ is at least m. We prove the existence of a shattered mistake tree of depth $\tilde{\Omega}(m)$. Like in Theorem 2, we prove this using the existence of a balanced point (see Equation (5)). Indeed, given a clique δ of size 2^m in $G_m(\mathcal{H})$, we pick the root of the mistake tree to be a balanced point x in δ, and proceed to find a balanced point in each of $\{S \in \delta : (x, 0) \in S\}$ and $\{S \in \delta : (x, 1) \in S\}$, corresponding to the left and right subtrees. We continue this way until one of the branches is consistent with exactly one dataset in δ. A standard calculation shows that this way one obtains a Littlestone tree of depth at least $\tilde{\Omega}(m)$. We note that there is a simple example of a class $\mathcal{H}$ that satisfies $\texttt{LD}(\mathcal{H}) < \texttt{CD}(\mathcal{H})$ (see Section 5). We leave as an open question to determine whether $\texttt{LD}(\mathcal{H}) = \Theta(\texttt{CD}(\mathcal{H}))$.

We now move to overview the proof of Theorem 6. Our proof follows by relating the representation dimension with the fractional clique dimension, showing that one is finite if and only if the other is finite. This finishes the proof because the representation dimension characterizes pure DP learnability (BNS19). The direction showing that finite representation dimension implies finite fractional clique dimension follows directly by results from (BNS19). In particular, (BNS19) proved a boosting result which reduces the probability of error of the random class from $1/4$ to ϵ. Applying their result with $\epsilon < 1/m$ yields a bound on the fractional clique number.

For the converse direction, we use Theorem 3. Assume that the fractional clique dimension of $\mathcal{H}$ is finite; therefore, for every m the fractional clique number is bounded by $\mathrm{poly}(m)$. Thus, by LP duality there exists a fractional coloring c with $\mathrm{poly}(m)$ many colors. Now, since fractional colorings correspond to distributions μ over hypotheses satisfying Equation (1), we can define a distribution over hypothesis classes by sampling $\mathrm{poly}(m)$ independent hypotheses from c. By Equation (1) it follows that for every realizable dataset S of size m, one of the $\mathrm{poly}(m)$ hypotheses is consistent with S with probability at least a constant (say $1/4$). By a standard generalization argument, this yields the desired bound on the representation dimension.

5. Future Work and Open Questions

Clique Dimension vs. Littlestone Dimension. Lemmas 16 and 20 tie together the clique dimension and the Littlestone dimension of a class $\mathcal{H}$, showing that they are equivalent up to log factors. It is natural to ask whether a tighter relationship holds:

Question 14 *Is it the case that* $\mathrm{CD}(\mathcal{H}) = \Theta(\mathrm{LD}(\mathcal{H}))$?

Notice that $\mathrm{LD}(\mathcal{H}) \leq \mathrm{CD}(\mathcal{H})$ for every $\mathcal{H}$, because the datasets corresponding to the branches of a shattered tree form a clique. Thus, it suffices to determine whether $\mathrm{CD}(\mathcal{H}) \leq O(\mathrm{LD}(\mathcal{H}))$ in order to answer the above question. We remark that there are classes $\mathcal{H}$ for which $\mathrm{LD}(\mathcal{H}) < \mathrm{CD}(\mathcal{H})$; for example, the following class $\mathcal{H} \subseteq \{0,1\}^4$ has $\mathrm{CD}(\mathcal{H}) = 3$ and $\mathrm{LD}(\mathcal{H}) = 2$:

$$\mathcal{H} = \big\{(0,0,0,1),(0,1,1,0),(0,1,1,1),(1,0,1,0), \qquad \text{(The red datasets form a clique.)}$$
$$(1,0,0,1),(1,1,1,0),(1,1,1,1),(1,1,0,1)\big\}.$$

Sauer-Shelah-Perles Lemma for Fractional Clique Dimension. In Appendix B we proved the analogue to the Sauer-Shelah-Perles Lemma for the fractional clique dimension. We showed a polynomial bound on the fractional clique number of $G_m(\mathcal{H})$ whenever $\mathrm{CD}^\star(\mathcal{H})$ is finite. In contrast to the polynomial-exponential dichotomy satisfied by the clique dimension, where the dimension itself bounds the polynomial degree, in this case the obtained bound on the polynomial degree depends on the difference $2^m - \omega_m^\star$ where m is any number satisfying $2^m - \omega_m^\star > 0$.

Question 15 *Let* $\mathcal{H}$ *be a class with* $\mathrm{CD}^\star(\mathcal{H}) = d < \infty$. *Is there a polynomial* $P_d(m)$ *whose degree depends only on* d *such that* $\omega_m^\star(\mathcal{H}) \leq P_d(m)$ *for every* m?

Direct Proofs. It will be interesting to find direct proofs for the characterizations of pure and approximate private learnability via the clique and fractional clique dimensions. In particular, is there a natural way to construct hard distributions for private learning from large cliques or fractional cliques?

Expressivity of Contradiction Graphs. Which learning theoretic properties are definable by the contradiction graph? In this work, we demonstrated that online learnability, approximate private learnability, and pure private learnability are captured by the clique and fractional clique numbers of the contradiction graph. How about PAC learnability (which is equivalent to finite VC dimension)? Is the property of having a finite VC dimension detectable in the contradiction graph? More formally, are there two hypothesis classes $\mathcal{H}_1, \mathcal{H}_2$ such that the VC dimension of $\mathcal{H}_1$ is finite and the VC dimension of $\mathcal{H}_2$ is infinite, but $G_m(\mathcal{H}_1) \equiv G_m(\mathcal{H}_2)$ for every m? (Here "$\equiv$" denotes the isomorphism relation between undirected graphs).

Acknowledgements

References

[ABL+22]   Noga Alon, Mark Bun, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private and online learnability are equivalent. *J. ACM*, 69(4):28:1–28:34, 2022.

[ALMM19]   Noga Alon, Roi Livni, Maryanthe Malliaris, and Shay Moran. Private PAC learning implies finite Littlestone dimension. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pages 852–860, New York, NY, USA, 2019. Association for Computing Machinery.

[app16a]   Apple promises to deliver AI smarts without sacrificing your privacy. *The Verge*, 2016.

[app16b]   Apple tries to peek at user habits without violating privacy. *The Wall Street Journal*, 2016.

[BDPSS09]   Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

[BGH+23]   Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization, 2023.

[BLM20]   Mark Bun, Roi Livni, and Shay Moran. An equivalence between private classification and online prediction. In Sandy Irani, editor, *61st IEEE Annual Symposium on Foun-*

dations of Computer Science, FOCS 2020, Durham, NC, USA, November 16-19, 2020, pages 389–402. IEEE, 2020.

[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of private learners. In *ITCS*. ACM, 2013.

[BNS19] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Characterizing the sample complexity of pure private learners. *Journal of Machine Learning Research*, 20(146):1–33, 2019.

[DLS$^+$] Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanava-jjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Lelerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd. The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. Association for Computing Machinery.

[FX15] Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. *SIAM Journal on Computing*, 44(6):1740–1764, 2015.

[ILPS22] Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 818–831. ACM, 2022.

[Kak43] Shizuo Kakutani. Notes on infinite product measure spaces. I, II. *Proc. Imp. Acad. Tokyo*, 19:148–151, 184–188, 1943.

[Kel75] J.L. Kelley. *General Topology*. Graduate Texts in Mathematics. Springer New York, 1975.

[Lit88] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[LM20] Roi Livni and Shay Moran. A limitation of the pac-bayes framework. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual*

Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020.

[MM22] Maryanthe Malliaris and Shay Moran. The unstable formula theorem revisited. *CoRR*, abs/2212.05050, 2022.

[PNG22] Aditya Pradeep, Ido Nachum, and Michael Gastpar. Finite littlestone dimension implies finite information complexity. In *IEEE International Symposium on Information Theory, ISIT 2022, Espoo, Finland, June 26 - July 1, 2022*, pages 3055–3060. IEEE, 2022.

[Rud87] W. Rudin. *Real and Complex Analysis*. Mathematics series. McGraw-Hill, 1987.

[Rud91] W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991.

[Sau72] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[Sio58] Maurice Sion. On general minimax theorems. *Pac. J. Math.*, 8:171–176, 1958.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[SU13] Edward R. Scheinerman and Daniel H. Ullman. *Fractional Graph Theory: a Rational Approach to the Theory of Graphs*. Dover Publications, Minola, N.Y., 2013.

[Tao11] T. Tao. *An Introduction to Measure Theory*. Graduate studies in mathematics. American Mathematical Society, 2011.

[Vad17] Salil P. Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer International Publishing, 2017.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

[VC68] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Proc. USSR Acad. Sci.*, 181(4):781–783, 1968.

Appendix A. Approximate Privacy and Cliques

In this section, we prove Theorems 2 and 7. Recall, Theorem 2 states that the clique number of the contradiction graph obeys an exponential-polynomial dichotomy, and Theorem 7 states that a class is approximately DP PAC learnable if and only if it's clique dimension is finite. The main idea is to study the relations between the clique dimension and the Littlestone dimension. It is known that a class $\mathcal{H}$ is approximately DP learnable if and only if the Littlestone dimension $\text{LD}(\mathcal{H})$ is finite (ALMM19; BLM20). Therefore, to prove Theorem 7 it suffices to show that the clique dimension $\text{CD}(\mathcal{H})$ is finite if and only if $\text{LD}(\mathcal{H})$ is finite. Recall that $\text{LD}(\mathcal{H})$ is equal to the depth of the deepest mistake tree that is shattered by $\mathcal{H}$. Observe that if a class $\mathcal{H}$ shatters a mistake tree of depth m, then the 2^m datasets corresponding to the leaves of the shattered tree form a clique in $G_m(\mathcal{H})$. This observation proves the following lemma.

Lemma 16 *Let $\mathcal{H}$ be a hypothesis class. Then $\text{LD}(\mathcal{H}) \leq \text{CD}(\mathcal{H})$.*

In simple words, deep shattered trees imply large cliques. We will show that the opposite statement holds as well. The following lemma is the crux of our proof. It asserts that for any clique in $G_m(\mathcal{H})$ there exists an unlabeled data point x which *separates* a non-negligible fraction of the datasets in the clique.

Lemma 17 *Let $\mathcal{H}$ be an hypothesis class and let $C \subseteq V_m(\mathcal{H})$ be a clique in $G_m(\mathcal{H})$. Then, there exists an unlabeled data point $x \in \mathcal{X}$ which is* balanced *in the following sense: at least $\frac{|C|-1}{2m}$ datasets in C contain the labeled example $(x, 1)$ and at least $\frac{|C|-1}{2m}$ datasets in C contain the labeled example $(x, 0)$.*

Before proving this lemma let us remark that quantitative improvements in the above lemma translate to improvements in Theorem 2 and to tighter bounds relating the Littlestone and clique dimensions. We elaborate on this in Section 5.

Proof [Proof of Lemma 17] Denote $c := |C|$. Given a labeled example (x, b) denote by $C_{(x,b)}$ the set of all datasets in C that contains (x, b):

$$C_{(x,b)} := \{S \in C \mid (x, b) \in S\}.$$

We perform the following iterative process to find a balanced example x:

> **Eliminate Unbalanced Examples in Clique:**
>
> As long as there is $S \in C$ and $(x, b) \in S$ such that $|C_{(x,1-b)}| < \frac{c-1}{2m}$, do:
> 1. Update $S \to S \smallsetminus \{(x, b)\}$.
> 2. Update the edges of C accordingly: if a dataset $S' \in C_{(x,1-b)}$ does not contradict S anymore (i.e. S and S' disagreed only x), then delete the edge between them.

Observe that:
1. The number of iterations is at most cm: Consider the sum $\sum_{S \in C} |S|$. At the beginning of the process $\sum_{S \in C} |S| = cm$, and at each iteration, the size of one dataset is reduced by one and hence $\sum_{S \in C} |S|$ decreases by one. The bound follows since $\sum_{S \in C} |S|$ is always non-negative.
2. At each iteration, the number of edges that are deleted is less than $\frac{c-1}{2m}$.

3. Thus, the total number of edges that are deleted during the process is less than $cm \cdot \frac{c-1}{2m} = \binom{c}{2}$. Therefore, at the end of the process, there is at least one remaining edge. That is, there are two datasets in C that contradict each other on an unlabeled example x which satisfies both $|C_{(x,1)}|$ and $|C_{(x,0)}|$ are at least $\frac{c-1}{2m}$, as required. ∎

Lemma 18 *Let $\mathcal{H}$ be a hypothesis class and denote by ω_m the clique number of $G_m(\mathcal{H})$. Then, for all m,*

$$\omega_m \leq (2m+1)^{\mathrm{LD}(\mathcal{H})} \leq (2m+1)^{\mathrm{CD}(\mathcal{H})}.$$

The idea of the proof is to apply Lemma 17 on a maximal clique in $G_m(\mathcal{H})$ and construct inductively a shattered Littlestone tree whose depth is large (as a function of ω_m). Together with the fact that the depth is at most $\mathrm{LD}(\mathcal{H})$, the desired bound follows.

Note that proving this lemma completes the proof of Theorem 2: if $\mathrm{LD}(\mathcal{H}) = \infty$ then by previous observation $\omega_m = 2^m$ for every m, and if $\mathrm{LD}(\mathcal{H})$ is finite then by the above lemma $\omega_m \leq (2m+1)^{\mathrm{LD}(\mathcal{H})} = \mathtt{poly}(m)$ for every m.

Proof [Proof of Lemma 18] Let C be a clique of size ω_m in $G_m(\mathcal{H})$. Without loss of generality, we can assume $\omega_m > 2m$ (otherwise the bound trivially holds[3]). By Lemma 17 there exists an unlabeled example x such that each of the sets

$$R = \{S \in C \mid (x,1) \in S\},$$
$$L = \{S \in C \mid (x,0) \in S\}$$

has size at least $\frac{\omega_m - 1}{2m} \geq \frac{\omega_m}{2m+1}$. Take x to be the root of a mistake tree, and recursively repeat the same operation on the sub-cliques induced by R, L. This way, in the i'th step we have 2^i cliques and the size of each clique is at least

$$\frac{\omega_m}{(2m+1)^i}.$$

Say this process terminates after T steps, yielding a shattered tree of depth $T \leq \mathrm{LD}(\mathcal{H})$. Note that the process terminates if and only if at least one of the produced cliques has size ≤ 1, hence

$$\frac{\omega_m}{(2m+1)^T} \leq 1,$$

which implies

$$\omega_m \leq (2m+1)^T \leq (2m+1)^{\mathrm{LD}(\mathcal{H})}.$$

The second inequality holds by Lemma 16. ∎

As an immediate corollary we conclude that for every class $\mathcal{H}$, the clique dimension is finite if and only if the Littlestone dimension is finite.

Corollary 19 *Let $\mathcal{H}$ be a hypothesis class. Then*

$$\mathrm{CD}(\mathcal{H}) < \infty \iff \mathrm{LD}(\mathcal{H}) < \infty$$

As shown earlier, this completes the proof of Theorem 7. We turn to state another quantitative relation between the Littlestone and clique dimensions that follows from Lemma 18.

---

3. If $\omega_m \leq 2m$ and $\mathrm{LD}(\mathcal{H}) \geq 1$, then the inequality in the lemma holds. In the degenerate case when $\mathrm{LD}(\mathcal{H}) = 0$, we have $|\mathcal{H}| = 1$ hence $\omega_m = 1$ for all m, and the inequality holds as well.

Lemma 20 *Let $\mathcal{H}$ be a hypothesis class. Then*

$$\mathtt{CD}(\mathcal{H}) \leq \max\big\{2\mathtt{LD}(\mathcal{H})\log(\mathtt{LD}(\mathcal{H})), 300\big\}.$$

The proof of this lemma is straightforward and technical thus deferred to Appendix D.

Appendix B. Pure Privacy and Fractional Cliques

In this section we prove Theorems 3 and 6. Recall, Theorem 3 states that the fractional clique number of the contradiction graph obeys an exponential-polynomial dichotomy, and Theorem 6 states that a class is pure DP PAC learnable if and only if it's fractional clique dimension is finite. Beimel et al. proved that a class $\mathcal{H}$ is pure DP learnable if and only if the representation dimension $\mathtt{RepDim}(\mathcal{H})$ is finite (BNS13). Recall, $\mathtt{RepDim}(\mathcal{H})$ is finite if there exists a distribution over finite hypothesis classes such that for every realizable distribution over labeled examples, with high probability, a random class contains a hypothesis that has small population error. We will show that $\mathtt{RepDim}(\mathcal{H})$ is finite if and only if the fractional clique dimension $\mathtt{CD}^\star(\mathcal{H})$ is finite. We begin with a technical lemma. An optimal fractional coloring of $G_m(\mathcal{H})$ induces a distribution over hypotheses such that every realizable dataset of size m is consistent with a random hypothesis with probability of at least $1/\omega_m^\star$ (Corollary 13). In other words, the probability that a random hypothesis has zero empirical loss (with respect to any realizable dataset of size m) is bounded from below by a positive constant. Using measure concentration arguments gives similar results when considering the population loss of a random hypothesis instead of the empirical loss. That is, there exists a distribution over hypotheses such that for every realizable distribution $\mathcal{D}$, the probability that a random hypothesis has small population loss (with respect to $\mathcal{D}$) is bounded from below. This fact is essential for the proof of the equivalence between the representation dimension and the fractional clique dimension.

Lemma 21 *Let $\mathcal{H}$ be a class and $m \in \mathbb{N}$. Then there exists a distribution $\mu^\star$ over hypotheses which satisfies the following. For every distribution $\mathcal{D}$ over labeled examples which is realizable by $\mathcal{H}$, and for every $0 \leq \theta \leq 1$,*

$$\Pr_{h \sim \mu^\star} [\mathtt{L}_\mathcal{D}(h) \leq \theta] \geq \frac{1}{\omega_m^\star} - (1 - \theta)^m,$$

where $\omega_m^\star$ is the fractional clique number of $G_m(\mathcal{H})$.

Proof The proof idea is to consider an optimal fractional coloring of $G_m(\mathcal{H})$ (which exists by Theorem 8). The number of colors of this coloring is the fractional chromatic number, which equals to the fractional clique number. Fractional colorings correspond to distributions over hypotheses, as demonstrated by Lemma 12. This optimal fractional coloring translates to an optimal distribution $\mu^\star$ over hypotheses. Then, we use a concentration argument to reason that this distribution achieves the desired bound.

Let $\mu^\star$ be, as in Corollary 13, a distribution over hypotheses which satisfied the following: for every realizable dataset S of size m,

$$\Pr_{h \sim \mu^\star} [h \text{ is consistent with } S] \geq \frac{1}{\omega_m^\star}.$$

Since this is true for every realizable dataset S of size m, it also holds on expectation when sampling S from a realizable distribution. Let $\mathcal{D}$ be a realizable distribution over labeled examples. Hence,

$$\frac{1}{\omega_m^\star} \leq \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \left[\mathop{\mathbb{E}}_{h \sim \mu^\star} \left[\mathbb{1}_{\{h \text{ is consistent with } S\}} | S \right] \right]$$

$$= \mathop{\mathbb{E}}_h \left[\mathop{\mathbb{E}}_S \left[\mathbb{1}_{\{h \text{ is consistent with } S\}} | h \right] \right]$$

$$= \mathop{\mathbb{E}}_h \left[(1 - \mathtt{L}_\mathcal{D}(h))^m \right].$$

Denote a random variable $X = \left(1 - \mathtt{L}_\mathcal{D}(h)\right)^m$, where $h \sim \mu^\star$. Thus, for every $\theta \in [0, 1]$,

$$\mathop{\mathbb{E}}_{h \sim \mu^\star} [X] \leq (1-\theta)^m \cdot \Pr\left[X < (1-\theta)^m\right] + 1 \cdot \Pr\left[X \geq (1-\theta)^m\right] \quad (X \leq 1 \text{ almost surely})$$

$$\leq (1-\theta)^m + \Pr\left[X \geq (1-\theta)^m\right]$$

$$= (1-\theta)^m + \mathop{\Pr}_{h \sim \mu^\star}\left[\mathtt{L}_\mathcal{D}(h) \leq \theta\right].$$

Hence,

$$\mathop{\Pr}_{h \sim \mu}\left[\mathtt{L}_\mathcal{D}(h) \leq \theta\right] \geq \mathbb{E}[X] - (1-\theta)^m$$

$$\geq \frac{1}{\omega_m^\star} - (1-\theta)^m.$$

∎

We now turn to prove Theorem 6. As elaborated above, it suffices to show that for every class $\mathcal{H}$, the dimension $\mathtt{RepDim}(\mathcal{H})$ is finite if and only if $\mathtt{CD}^\star(\mathcal{H})$ is finite. A key tool used in the proof is the SSP lemma for the fractional clique dimension, which states that $\mathtt{CD}^\star(\mathcal{H})$ is finite if and only if for every m, the fraction clique number of $G_m(\mathcal{H})$ is bounded by a polynomial in m (Theorem 3). We begin by proving Theorem 6 assuming Theorem 3, and then we will formally prove Theorem 3.

Lemma 22 (Finite fractional clique dimension → finite representation dimension) *Let $\mathcal{H}$ be an hypothesis class and assume $\mathtt{CD}^\star(\mathcal{H}) = d < \infty$. Denote $\epsilon = \frac{1}{\omega_{d+1}^\star} - \frac{1}{2^{d+1}}$ where $\omega_{d+1}^\star$ is the fractional chromatic number of $G_{d+1}(\mathcal{H})$. Then,*

$$\mathtt{RepDim}(\mathcal{H}) = O\left(\frac{\log \frac{1}{\epsilon}}{\epsilon^2} \cdot \ln\left(\frac{\log \frac{1}{\epsilon}}{\epsilon^2}\right) + \frac{\log \frac{1}{\epsilon}}{\epsilon^2} \cdot \ln \ln\left(\frac{\log \frac{1}{\epsilon}}{\epsilon^2}\right)\right).$$

Proof The goal is to construct a distribution $\mathcal{P}$ over finite hypothesis classes such that for every realizable distribution over labeled examples $\mathcal{D}$:

$$\mathop{\Pr}_{\mathcal{C} \sim \mathcal{P}}\left[(\exists h \in \mathcal{C}) : \mathtt{L}_\mathcal{D}(h) \leq \frac{1}{4}\right] \geq \frac{3}{4}. \tag{7}$$

By Theorem 3 there exists a natural number $\alpha = O\left(\log\left(\frac{1}{\epsilon}\right)/\epsilon^2\right)$ such that for every m, we have $\omega_m^\star \leq m^\alpha$. Let $m = m(\alpha)$ to be determined later on. By applying Lemma 21 with $\theta = \frac{1}{4}$, it follows that there exists a distribution $\mu^\star$ over hypotheses such that for every realizable distribution $\mathcal{D}$,

$$\mathop{\Pr}_{h \sim \mu^\star}\left[\mathtt{L}_\mathcal{D}(h) \leq \frac{1}{4}\right] \geq \frac{1}{m^\alpha} - \left(\frac{3}{4}\right)^m =: q(m). \tag{8}$$

19

Sampling k hypotheses $h_1, \ldots, h_k$ i.i.d. from $\mu^\star$ induces a probability distribution over classes of size at most k. Denote this distribution by $\mathcal{P}_k$. Observe that

$$\Pr_{\mathcal{C} \sim \mathcal{P}_k} \left[\exists h \in \mathcal{C}, L_{\mathcal{D}}(h) \leq \frac{1}{4} \right] = 1 - \Pr_{\{h_i\}_{i=1}^k \sim (\mu^\star)^k} \left[(\forall i) : L_{\mathcal{D}}(h_i) > \frac{1}{4} \right]$$

$$\geq 1 - \left(1 - q(m) \right)^k. \qquad \text{(by Equation (8))}$$

The following technical lemma concludes the proof:

Lemma 23 *Let $\alpha \geq 2$, and set $m = \lfloor 20\alpha \ln \alpha \rfloor$ and $k = 4m^\alpha$. Then,*

$$\left(1 - q(m) \right)^k \leq \frac{1}{4}.$$

The proof of Lemma 23 is deferred Appendix D. As a result, by setting m, k as in Lemma 23, the distribution $\mathcal{P}_k$ satisfies the property described in Equation (7) and therefore $\texttt{RepDim}(\mathcal{H}) = O(\ln k) = O(\alpha \ln \alpha + \alpha \ln \ln \alpha) < \infty$.

∎

Lemma 24 (Finite representation dimension → finite fractional clique dimension) *Let $\mathcal{H}$ be a class and assume $\texttt{RepDim}(\mathcal{H}) = d < \infty$. Then $\texttt{CD}^\star(\mathcal{H}) < \infty$.*

Proof We need to show that there exists a natural number m such that the fractional clique number of $G_m(\mathcal{H})$ is strictly smaller than 2^m. By Lemma 12 it is enough to show that there exist m, a distribution μ over hypotheses, and $\epsilon > 0$ such that for every realizable dataset S of size m,

$$\Pr_{h \sim \mu} [h \text{ is consistent with } S] \geq \frac{1}{2^m} + \epsilon. \qquad (9)$$

By Lemma 10, there exists a distribution $\mathcal{P}$ over classes of size $k = O((2m)^{d + \ln \ln \ln(2m) + \ln \ln(4)})$ such that for every realizable distribution $\mathcal{D}$,

$$\Pr_{\mathcal{C} \sim \mathcal{P}} \left[\exists h \in \mathcal{C} \text{ s.t. } L_{\mathcal{D}}(h) \leq \frac{1}{2m} \right] \geq \frac{3}{4}. \qquad (10)$$

Randomly sampling $\mathcal{C} \sim \mathcal{P}$ and then randomly uniformly sampling $h \in \mathcal{C}$ induces a probability distribution over hypotheses. Denote this distribution by μ. It suffices to show that μ satisfies the property in Equation (9). Let $S = ((x_1, y_1), \ldots, (x_m, y_m))$ be a realizable dataset of size m. Define a distribution over labeled examples $\mathcal{D}$ as follows

$$\mathcal{D}(x, y) = \frac{1}{m} \cdot \sum_{i=1}^m \mathbb{1}[(x, y) = (x_i, y_i)].$$

Note that $\mathcal{D}$ is realizable with respect to $\mathcal{H}$ since S is a realizable dataset. Furthermore, for every hypothesis h,

$$L_{\mathcal{D}}(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}[h(x_i) \neq y_i] = L_S(h). \qquad (\text{by definition of } \mathcal{D})$$

Therefore, h is consistent with S if and only if $\text{L}_\mathcal{D}(h) < \frac{1}{m}$. Therefore,

$$
\begin{aligned}
\Pr_{h \sim \mu}\left[h \text{ is consistent with } S\right] &\geq \Pr_{h \sim \mu}\left[\text{L}_\mathcal{D}(h) \leq \frac{1}{2m}\right] \\
&\geq \Pr_{\mathcal{C} \sim \mathcal{P}}\left[\exists h \in \mathcal{C} \text{ s.t. } L_\mathcal{D}(h) \leq \frac{1}{2m}\right] \cdot \frac{1}{k} \qquad \text{(by definition of } \mu) \\
&\geq \frac{3}{4k} \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{(by Equation (10))} \\
&\geq \frac{1}{2^m} + \epsilon,
\end{aligned}
$$

where the last inequality holds with some $\epsilon > 0$ provided that $m > \log(\frac{4k}{3})$. Thus, it remains to show that there exists m such that $m > \log(\frac{4k}{3})$. (Recall that $k = k(m)$ is the size of the hypothesis classes in the support of the distribution $\mathcal{P}$ such that Equation (10) holds.) Now, since $k = O\left((2m)^{d + \ln \ln \ln(2m) + \ln \ln(4)}\right)$ (Lemma 10), we get that $\log(\frac{4k}{3})$ is sublinear in m. Thus, $m \geq \log(\frac{4k}{3})$ for a large enough m, as required. ∎

B.1. SSP Lemma for Fractional Clique Dimension

We turn to prove Theorem 3. We begin with a general layout of the proof. Let $\mathcal{H}$ be a class and assume that Item 1 does not hold. Meaning, there exists a natural number m_0 such that the fractional chromatic number of $G_{m_0}(\mathcal{H})$ is strictly smaller than 2^{m_0}. In order to show that $\omega_m^\star$ is bounded by a polynomial, by Lemma 12 it is enough to show that for every m, there exists a distribution over hypotheses μ_m, such that for every realizable dataset S of size m, a random hypothesis is consistent with S with probability at least $\texttt{poly}(m^{-1})$; i.e. for every realizable dataset of size m,

$$
\Pr_{h \sim \mu_m}\left[h \text{ is consistent with } S\right] \geq \frac{1}{\texttt{poly}(m)}. \tag{11}
$$

Specifically, we will derive an explicit upper bound on the degree of the polynomial on the right hand side, denoted by α:

$$
\alpha = O\left(\frac{log\frac{1}{\epsilon}}{\epsilon^2}\right), \tag{12}
$$

where $\epsilon = \frac{1}{\omega_{m_0}^\star} - \frac{1}{2^{m_0}}$. The idea is to use a boosting argument to show such distributions exist. The analysis involves a reduction to regret analysis for online predictions using experts' advice.

Step 1: There exists a distribution $\tilde{\mu}$ over hypotheses such that a random hypothesis is "slightly better then a random guess". Formally, for every realizable distribution over labeled examples $\mathcal{D}$, and for every $\gamma \in (0, \frac{1}{2})$,

$$
\Pr_{h \sim \tilde{\mu}}[\text{L}_\mathcal{D}(h) \leq \frac{1}{2} - \gamma] > \epsilon - 2\gamma.
$$

Step 2: Set $T = \lceil \frac{2 \log m}{\gamma^2} \rceil$. For every realizable dataset S of size m, the majority vote of T i.i.d. hypotheses sampled from $\tilde{\mu}$ (from Step 1) is consistent with S, with probability at least $(\epsilon - 2\gamma)^T = m^{-\frac{2}{\gamma^2} \log \frac{1}{\epsilon - 2\gamma}}$.

Step 3: Denote by μ_m the a distribution over hypotheses induced by sampling T i.i.d. hypotheses from $\tilde{\mu}$ and then taking their majority vote. Plug in $\gamma = \frac{\epsilon}{4}$ and conclude

$$\Pr_{h \sim \mu_m} [h \text{ is consistent with } S] \geq m^{-O\left(\frac{\log \frac{1}{\epsilon}}{\epsilon^2}\right)},$$

which concludes the proof.

Proof [Proof of Theorem 3] From Lemma 21 there exists a distribution over hypotheses $\tilde{\mu}$ such that for every realizable distribution $\mathcal{D}$

$$\Pr_{h \sim \tilde{\mu}} \left[\mathtt{L}_{\mathcal{D}}(h) \leq \frac{1}{2} - \gamma \right] \geq \frac{1}{\omega_{m_0}^{\star}} - \left(\frac{1}{2} + \gamma \right)^{m_0}$$

$$= \frac{1}{2^{m_0}} + \epsilon - \left(\frac{1}{2} + \gamma \right)^{m_0}$$

where $0 < \gamma < \frac{1}{2}$ will be determined later on. Observe that since $\left(\frac{1}{2} + \gamma\right)^{m_0} < \frac{1}{2^{m_0}} + 2\gamma$ for all $0 < \gamma < \frac{1}{2}$[4], we get that

$$\Pr_{h \sim \tilde{\mu}} \left[\mathtt{L}_{\mathcal{D}}(h) \leq \frac{1}{2} - \gamma \right] \geq \epsilon - 2\gamma. \tag{13}$$

Here and below, we say that an hypothesis h is γ-*good* with respect to a distribution $\mathcal{D}$ if $\mathtt{L}_{\mathcal{D}}(h) \leq \frac{1}{2} - \gamma$. Let m be a natural number and $S = \{(x_i, y_i)\}_{i=1}^m$ be a realizable dataset of size m. Define $\mu = \mu_m$ to be the distribution over hypotheses induced by taking the majority vote of T i.i.d. hypotheses sampled from $\tilde{\mu}$, where $T = \lceil \frac{2 \log m}{\gamma^2} \rceil$.

The goal is to prove that the distribution μ satisfies the property described in Equation (11). The following two lemmas will complete the proof.

Lemma 25 *Let S be a realizable dataset of size m and $h_1, \ldots, h_T$ a sequence of hypotheses where $T = \lceil \frac{2 \log m}{\gamma^2} \rceil$. Then, there exists a sequence of realizable distributions $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T$ such that $\mathcal{D}_t$ is a function of $h_1, \ldots, h_{t-1}$ and the following condition holds. If for every $t \in [T]$, h_t is γ-good with respect to $\mathcal{D}_t$, then S is consistent with $\mathrm{MAJ}\{h_t\}_{t=1}^T$.*

The proof of Lemma 25 uses a regret analysis for online learning using expert advice, and is deferred after a short discussion introducing concepts and notations from online learning.

Lemma 26 *For every $t \leq T$,*

$$\Pr_{(h_1, \ldots, h_t) \sim \tilde{\mu}^t} [h_k \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_k, \forall k = 1, \ldots, t] \geq (\epsilon - 2\gamma)^t,$$

where $\mathcal{D}_1, \ldots, \mathcal{D}_t$ is a sequence of realizable distributions as in Lemma 25.

The proof of Lemma 26 is technical and follows from Equation (13) and simple induction.

---

4. $\left(\frac{1}{2} + \gamma\right)^{m_0} = \frac{1}{2^{m_0}} + \gamma \sum_{k=0}^{m_0-1} \binom{m_0}{k} \gamma^{m_0-1-k} \cdot \frac{1}{2^k} < \frac{1}{2^{m_0}} + \gamma \frac{1}{2^{m_0-1}} \sum_{k=0}^{m_0-1} \binom{m_0}{k} = \frac{1}{2^{m_0}} + \gamma \frac{2^{m_0}-1}{2^{(m_0-1)}} < \frac{1}{2^{m_0}} + 2\gamma$

With Lemma 25 and Lemma 26 in hand, the following calculation finishes the proof.

$$\Pr_{h\sim\mu}\left[h \text{ is consistent with } S\right] = \Pr_{(h_1,\ldots,h_T)\sim\tilde{\mu}^T}\left[S \text{ is consistent with } \mathrm{MAJ}\{h_t\}_{t=1}^{T}\right]$$

$$\text{(by the definition of } \mu)$$

$$\geq \Pr_{(h_1,\ldots,h_T)\sim\tilde{\mu}^T}\left[h_t \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_t, \forall t = 1,\ldots T\right]$$

$$\text{(from Lemma 25)}$$

$$\geq (\epsilon - 2\gamma)^T \qquad\qquad \text{(from Lemma 26)}$$

$$\geq (\epsilon - 2\gamma)^{\frac{2\log m}{\gamma^2}} \qquad\qquad \text{(by the choice of } T)$$

$$= m^{-\frac{2}{\gamma^2}\log\frac{1}{\epsilon-2\gamma}}.$$

Denote $\alpha(\gamma) = \frac{2}{\gamma^2}\log\frac{1}{\epsilon-2\gamma}$. The upper bound stated in Equation (12) is obtained by plugging in $\gamma = \frac{\epsilon}{4}$. $\blacksquare$

Proof [Proof of Lemma 26] We will show that for $t \leq T$,

$$\Pr_{(h_1,\ldots,h_t)\sim\tilde{\mu}^t}\left[h_k \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_k, k = 1,\ldots,t\right] \geq (\epsilon - 2\gamma)^t. \tag{14}$$

Indeed, the base case where $t = 1$ follows directly from Equation (13) and the fact that $\mathcal{D}_1$ does not depend on h_1. For $t > 1$,

$$\Pr_{(h_1,\ldots,h_t)\sim\tilde{\mu}^t}\left[h_k \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_k, k = 1,\ldots,t\right] =$$

$$= \mathbb{E}_{(h_1,\ldots,h_t)\sim\tilde{\mu}^t}\left[\mathbb{1}_{\{h_k \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_k, k=1,\ldots,t-1\}} \cdot \mathbb{1}_{\{h_t \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_t\}}\right]$$

$$= \mathbb{E}_{h_1,\ldots,h_{t-1}}\left[\mathbb{E}_{h_t}\left[\mathbb{1}_{\{h_k \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_k, k=1,\ldots,t-1\}} \cdot \mathbb{1}_{\{h_t \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_t\}} \mid h_1,\ldots,h_{t-1}\right]\right] \tag{15}$$

$$= \mathbb{E}_{h_1,\ldots,h_{t-1}}\left[\mathbb{1}_{\{h_k \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_k, k=1,\ldots,t-1\}} \cdot \mathbb{E}_{h_t}\left[\mathbb{1}_{\{h_t \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_t\}} \mid h_1,\ldots,h_{t-1}\right]\right] \tag{16}$$

$$\geq (\epsilon - 2\gamma) \mathbb{E}_{h_1,\ldots,h_{t-1}}\left[\mathbb{1}_{\{h_k \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_k, k=1,\ldots,t-1\}}\right] \tag{17}$$

$$\geq (\epsilon - 2\gamma)^t. \tag{18}$$

Equation (15) is obtained by applying the law of total expectation. Equation (16) holds since for every $k < t$, $\mathcal{D}_k$ does not depend on h_t. Equation (17) holds since for every realization of $h_1,\ldots,h_{t-1}$, $\mathcal{D}_t$ is determined and does not depend on h_t, hence by Equation (13),

$$\mathbb{E}_{h_t\sim\tilde{\mu}}\left[\mathbb{1}_{\{h_t \text{ is } \gamma\text{-good w.r.t. } \mathcal{D}_t\}} \mid h_1 = \tilde{h}_1,\ldots,h_{t-1} = \tilde{h}_{t-1}\right] \geq \epsilon - 2\gamma.$$

Finally, Equation (18) is true by induction. $\blacksquare$

In order to finish the proof of Theorem 3 it is left to prove Lemma 25. We begin with a short technical overview of online learning using experts' advice.

Learning Using Expert Advice. We briefly introduce the setting of online prediction using expert advice. Let $Z = \{z_1,\ldots,z_m\}$ be a set of experts and I be a set of instances. Given an instance $i \in I$, an expert z predicts a prediction $p = z(i) \in \{0,1\}$.

In this online learning problem, at each round t the learner receives an instance i_t and has to come up with a prediction according to advice from the m experts. The learner does that by obtaining at each round a distribution w_t over the set of experts Z, randomly picking an expert $z^{(t)} \sim w_t$, and predicting $p_t = z^{(t)}(i_t)$. Then, the learner receives f_t, the true label of i_t, and pays a cost $l_t = \mathbb{1}[p_t \neq f_t]$ for taking the advise of $z^{(t)}$. Lastly, the learner updates the distribution over the set of experts Z (according to some update rule).

Online predictions using expert advice-general scheme

Input: $Z = \{z_1, \ldots, z_m\}$ a set of experts.
Initialize: distribution w_1 over Z.
For every round $t = 1, 2, \ldots, T$:
 – Receive an instance i_t.
 – Sample an expert $z^{(t)} \sim w_t$.
 – Predict $p_t = z^{(t)}(i_t)$.
 – Receive true label f_t and pay cost of $l_t = \mathbb{1}[p_t \neq f_t]$.
 – Update the distribution w_t to w_{t+1}.

The *total loss* of the learner is the expected sum of costs:

$$L(T) = \mathbb{E}\left[\sum_{t=1}^{T} l_t\right].$$

The *regret* of the learner is the difference between the total loss and the loss of the best expert:

$$\mathsf{Reg}(T) = L(T) - \min_{z_j \in Z} \sum_{t=1}^{T} \mathbb{1}[z_j(i_t) \neq f_t].$$

The goal of the learner is to compete with the best expert in Z; i.e. the goal is to minimize the regret. There are several well studied algorithms for online prediction using expert advice which achive sublinear regret. A classic example is *Multiplicative Weights* (see section 21.2 in (SSBD14)) which satisfy the following regret bound:

$$\mathsf{Reg}(T) \leq \sqrt{2T \log m}. \tag{19}$$

We will use that fact to prove Lemma 25.

Proof [Proof of Lemma 25] Let $S = \{z_j = (x_j, y_j)\}_{j=1}^{m}$ be a realizable dataset of size m and let $h_1, \ldots, h_T$ be a sequence of hypotheses where $T = \lceil \frac{2 \log m}{\gamma^2} \rceil$. Recall, an hypothesis h is γ-good with respect to a distribution $\mathcal{D}$ if $\mathsf{L}_{\mathcal{D}}(h) \leq \frac{1}{2} - \gamma$. The goal is to show that there exists a sequence of realizable distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ which satisfies the following properties:

1. $\mathcal{D}_t = \mathcal{D}_t(h_1, \ldots, h_{t-1})$ for all $t \leq T$. In particular, $\mathcal{D}_1$ does not depend on $h_1, \ldots, h_T$.
2. If for every t, h_t is γ-good with respect to $\mathcal{D}_t$, then S is consistent with the majority hypothesis $\mathrm{MAJ}\{h_t\}_{t=1}^{T}$.

The idea of the proof is as follows. We think of the dataset S as a set of m "experts". Then we simulate an online learner for learning using experts' advice with $h_1, \ldots, h_T$ as "instances". The distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ will be induced by the internal weights of the "experts" over time during the simulation. Regret analysis, together with the specific choice of number of rounds T, will lead to the desired results.

Let $\mathcal{A}$ be an online learner for prediction using expert advice, which satisfies the regret bound in Equation (19), with the following settings: Our set of experts is the dataset $S = \{z_1, z_2, \ldots, z_m\}$, and a possible instance is an hypothesis $h \in \{0,1\}^{\mathcal{X}}$. Given an expert $z = (x,y)$ and an instance h, the loss of z on h is defined to be $l(z,h) = \mathbb{1}[h(x) = y]$; i.e. $\mathcal{A}$ suffers a loss if the hypothesis h is consistent with the example $z = (x,y)$. Now, when running $\mathcal{A}$ on the input sequence $h_1, \ldots, h_T$, the algorithm maintains a distribution w_t on the dataset S at each timestamp $t \in [T]$. Denote $\mathcal{D}_t$ to be a distribution over all labeled examples induced by w_t; i.e,

$$\mathcal{D}_t(x, y) = \sum_{j : z_j = (x,y)} w_t(z_j)$$

Note that since S is a realizable dataset, $\mathcal{D}_t$ is a realizable distribution for all t.

We claim that the sequence $\mathcal{D}_1, \ldots, \mathcal{D}_T$ fulfills the desired properties. Indeed, property (1) holds since the internal distribution obtained by the learner, w_t, does not depend on the current instance h_t. We next show that property (2) holds as well. assume h_t is γ-good with respect to $\mathcal{D}_t$ for all t. Then, the expected loss of $\mathcal{A}$ at timestamp t satisfies

$$\begin{aligned}
\mathbb{E}[l_t] &= \Pr_{(x,y) \sim \mathcal{D}_t}[h_t(x) = y] \\
&= 1 - L_{\mathcal{D}_t}(h_t) \\
&\geq \frac{1}{2} + \gamma.
\end{aligned}$$

By linearity, the expected total loss of $\mathcal{A}$ is bounded from below by

$$\mathbb{E}\left[\sum_{t=1}^{T} l_t\right] \geq \frac{T}{2} + \gamma T. \tag{20}$$

If by contradiction S is not consistent with $\mathrm{MAJ}\{h_1, \ldots, h_T\}$, then there exists $j \in [m]$ such that

$$\sum_{t=1}^{T} \mathbb{1}[h_t(x_j) = y_j] = \sum_{t=1}^{T} l(z_j, h_t) < \frac{T}{2}. \tag{21}$$

From the regret bound in Equation (19), combining together Equations (20) and (21) yields

$$\gamma T < \sqrt{2T \log m}.$$

However, this is a contradiction to the choice of T (recall $T = \lceil \frac{2 \log m}{\gamma^2} \rceil$). Thus property (2) indeed holds and the proof of the lemma is complete. ■

Appendix C. Strong Duality in the Contradiction Graph

We now turn to prove that for every (possibly infinite) $\mathcal{H}$ and $m \in \mathbb{N}$, the contradiction graph $G_m(\mathcal{H})$ satisfies that its fractional chromatic and clique numbers are equal and bounded by 2^m (Theorem 8). Towards this end it will be convenient to represent fractional colorings using probability distributions over $\{0,1\}^{\mathcal{X}}$ and fractional cliques using probability distributions over $V_m(\mathcal{H})$ – the set of all m-datasets that are realizable by $\mathcal{H}$.

Since $\mathcal{H}$ might be infinite, we need to be more careful with specifying which distributions we consider. Let us begin with $\{0,1\}^{\mathcal{X}}$: one of the basic facts we use in the paper is that the fractional chromatic number of $G_m(\mathcal{H})$ is at most 2^m. To prove this, we drew a random hypothesis by sampling its values on each $x \in \mathcal{X}$ uniformly and independently. For this argument to apply to

infinite domains $\mathcal{X}$ it is natural to use the product topology on $\{0,1\}^{\mathcal{X}}$ (i.e. the Tychonoff product of the discrete topology over $\{0,1\}$). The product structure allows to sample a random hypothesis as above, by taking the product of the uniform distributions over $\{0,1\}$ (see Remark 30).

Formally, we let the set of fractional colorings be the space of all Borel[5] regular[6] probability measures over $\{0,1\}^{\mathcal{X}}$, where the latter is equipped with the product topology (Tychonoff). Denote this space by $\Delta\big(\{0,1\}^{\mathcal{X}}\big)$. (The relevant concepts in topology are defined below.)

We now turn to define fractional cliques, which are distributions over the set of realizable datasets of size m, $V_m(\mathcal{H})$. Here, we simply consider finitely supported distributions and use the trivial discrete topology and the corresponding Borel σ-algebra over $V_m(\mathcal{H})$, both equal to the entire powerset. Denote the space of all finitely supported probability distributions over $V_m(\mathcal{H})$ by $\Delta(V_m^{\mathcal{H}})$.

We next extend the definitions of fractional clique and chromatic numbers using the above structures. The definitions rely on the following basic claim.

Claim 27 *The function* $f : V_m(\mathcal{H}) \times \{0,1\}^{\mathcal{X}} \to \mathbb{R}$ *defined by*

$$f(S, h) = \mathbb{1}[h \text{ is consistent with } S],$$

is continuous and hence Borel-measurable with respect to the product topology over $V_m(\mathcal{H}) \times \{0,1\}^{\mathcal{X}}$.

Proof Because open sets are union-closed, it is enough to show that for each fixed $S \in V_m(\mathcal{H})$, the set $\{(S, h) : h \text{ is consistent with } S\}$ is open. Since $\{S\}$ is open in the discrete topology on $V_m(\mathcal{H})$, it is enough to show that $\{h : h \text{ is consistent with } S\}$ is open in the product topology on $\{0,1\}^{\mathcal{X}}$. The latter is indeed true because this set is a basic open set in the product topology (see below for definitions and topological background). ∎

Definition 28 (Fractional Chromatic and Clique Numbers) *Let* $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ *be a class and* $m \in \mathbb{N}$. *The* fractional chromatic number *of* $G_m(\mathcal{H})$, *denoted* $\chi_m^{\star}$, *is defined by*

$$\frac{1}{\chi_m^{\star}} = \sup_{\mu \in \Delta\big(\{0,1\}^{\mathcal{X}}\big)} \inf_{\nu \in \Delta(V_m^{\mathcal{H}})} \mathbb{E}_{\substack{h \sim \mu, \\ S \sim \nu}} \big[\mathbb{1}[h \text{ is consistent with } S]\big],$$

with the convention that $\frac{1}{0} = \infty$ *and* $\frac{1}{\infty} = 0$.

The fractional clique number *of* $G_m(\mathcal{H})$, *denoted* $\omega_m^{\star}$, *is defined by*

$$\frac{1}{\omega_m^{\star}} = \inf_{\nu \in \Delta(V_m^{\mathcal{H}})} \sup_{\mu \in \Delta\big(\{0,1\}^{\mathcal{X}}\big)} \mathbb{E}_{\substack{h \sim \mu, \\ S \sim \nu}} \big[\mathbb{1}[h \text{ is consistent with } S]\big],$$

with the convention that $\frac{1}{0} = \infty$ *and* $\frac{1}{\infty} = 0$.

Remark 29 *The definition below of the fractional chromatic and clique numbers involves taking expectation of the random variable* $\mathbb{1}[h \text{ is consistent with } S]$ *when* $h \sim \mu \in \Delta\big(\{0,1\}^{\mathcal{X}}\big)$ *and* $S \sim$

---

5. A Borel measure on a topological space X is a measure defined on the Borel σ-algebra, which is the σ-algebra generated by all open sets in X.

6. Roughly speaking, a regular measure is a measure that can be approximated from above by open sets and from below by compact sets. See definition in Section C.1.

$\nu \in \Delta(V_m^{\mathcal{H}})$. *Note that because the distribution ν is finitely supported, this expectation can be expressed as a finite sum:*

$$\underset{\substack{h \sim \mu, \\ S \sim \nu}}{\mathbb{E}} \left[\mathbb{1}[h \text{ is consistent with } S] \right] = \sum \nu(S) \cdot \underset{h \sim \mu}{\mathbb{E}} \left[\mathbb{1}[h \text{ is consistent with } S] \right],$$

where the sum ranges over all (finitely many) datasets S in the support of ν.

Remark 30 *In Section C.3 we show that there exists a Borel-regular probability measure over $h \in \{0,1\}^{\mathcal{X}}$ which corresponds to sampling $h(x) \in \{0,1\}$ uniformly and independently for each $x \in \mathcal{X}$. Thus, the corresponding fractional coloring which witnesses that $\chi_m^{\star} \leq 2^m$ is well-defined.*

Sion's Theorem.   Our proof of Theorem 8 relies on Sion's theorem, which is a generalization of Von-Neumann's minimax theorem. Note that we state a slightly weaker version than the original Sion's theorem (Sio58).

Theorem 31 (Sion's Theorem – weak version) *Let W be a compact convex subset of a linear topological space[7] and U a convex subset of a linear topological space. If F is a real-valued function on $W \times U$ such that*
  1. *$F(w, \cdot)$ is linear and continuous on U for every $w \in W$, and*
  2. *$F(\cdot, u)$ is linear and continuous on W for every $u \in U$*
then,

$$\max_{w \in W} \inf_{u \in U} F(w, u) = \inf_{u \in U} \max_{w \in W} F(w, u).$$

To apply Sion's Theorem we start by defining appropriate topologies on $\Delta\left(\{0,1\}^{\mathcal{X}}\right)$ and $\Delta(V_m^{\mathcal{H}})$. We start by presenting useful facts from topology and functional analysis. Readers who are familiar with this material may skip it and continue to Section C.2

C.1. Preliminaries from Topology and Analysis

Product Topology.   Let $\{0,1\}^{\mathcal{X}}$ be the space of all functions $f : \mathcal{X} \to \{0,1\}$. The product topology on $\{0,1\}^{\mathcal{X}}$ is the coarsest topology in which each projection is continuous; i.e, for every $x \in \mathcal{X}$, the mapping $\pi_x : \{0,1\}^{\mathcal{X}} \to \{0,1\}$ defined by $\pi_x(f) = f(x)$, is continuous. A basis of open sets of the product topology is given by sets of the form

$$U_{(x_1,y_1)...,(x_m,y_m)} = \{f \mid f(x_i) = y_i \text{ for all } 1 \leq i \leq m\},$$

where $m \in \mathbb{N}$ and $\left((x_1, y_1), \ldots, (x_m, y_m)\right) \in \left(\mathcal{X} \times \{0,1\}\right)^m$. Note that every such basic open set is also closed since its complement is a union of $2^m - 1$ basic sets (such sets are often called *clopen* because they are both closed and open). By Tychonoff's theorem (e.g. see Theorem 5.13 in (Kel75)), the space $\{0,1\}^{\mathcal{X}}$ is compact for any set $\mathcal{X}$.

Total Variation.   We next state some basic useful facts on total variation. For more information see Chapter 6 at (Rud87). Let (X, Σ, μ) be a measure space where μ is a signed measure[8]. The

---

7. A linear topological space is a vector space that is also a topological space with the property that the vector space operations (vector addition and scalar multiplication) are continuous.

8. A signed measure is a σ-additive set-function that can assign negative values.

total variation measure of μ is a positive measure on (X, Σ) defined as follows

$$|\mu|(E) = \sup \sum_{i=1}^{\infty} |\mu(E_i)|, \quad E \in \Sigma$$

where the supremum ranges over all countable partitions $\{E_i\}_{i=1}^{\infty}$ of E. Note that for every measurable set E, $|\mu|(E) \geq |\mu(E)|$. If $|\mu|$ is finite, we say that μ has *bounded variation*, and the *total variation* of μ is defined to be

$$\|\mu\|_{TV} = |\mu|(X).$$

The set of all signed measures with bounded variation, together with $\|\cdot\|_{TV}$, is a normed linear space. Given a signed measure μ, define

$$\mu^+ = \frac{1}{2}(|\mu| + \mu), \quad \mu^- = \frac{1}{2}(|\mu| - \mu).$$

Then, both μ^+ and μ^- are positive measures on (X, Σ) which are called the *positive* and *negative variations* of μ. Also we have that $\mu = \mu^+ - \mu^-, |\mu| = \mu^+ + \mu^-$, and therefore

$$\|\mu\|_{TV} = |\mu|(X) = \mu^+(X) + \mu^-(X) = \|\mu^+\|_{TV} + \|\mu^-\|_{TV}.$$

This representation of μ as the difference of the positive measures μ^+ and μ^- is called the *Jordan Decomposition* of μ.

Regularity. A measure μ, defined on a $\sigma-$algebra Σ of a Hausdorff space X, is called *inner regular* if for every $E \in \Sigma$,

$$\mu(E) = \sup_{K \in \Sigma} \left(\mu(K) \mid K \subseteq E \text{ is compact} \right).$$

The measure μ is called *outer regular* if for every $E \in \Sigma$,

$$\mu(E) = \inf_{O \in \Sigma} \left(\mu(O) \mid O \supset E \text{ is open} \right).$$

The measure μ is *regular* if it is both inner regular and outer regular. Note that if X is compact and μ is finite, inner regularity is equivalent to outer regularity.

Weak* Topology and the Dual of $C(K)$. Recall that for a linear space V, the space $V^\star$ denotes the dual space, i.e. the space of all linear functionals on V. Let X be a linear topological space over $\mathbb{R}$. For $x \in X$, let $T_x \in X^{\star\star}$ denote the evaluation operator: $T_x(f) = f(x)$, where $f \in X^\star$ is a linear functional $f : X \to \mathbb{R}$. The *weak* topology* on $X^\star$ is the coarsest topology such that the operators T_x are continuous.

Let K be a compact Hausdorff space. Denote by $C(K)$ the set of all real-valued continuous functions on K, and denote by $\mathcal{B}(K)$ the set of all finite signed regular Borel measures on K. We treat both $C(K)$ and $\mathcal{B}(K)$ as linear normed spaces, the first is equipped with max-norm and the second with total-variation norm. By Riesz-Markov Representation Theorem $C(K)^\star$ and $\mathcal{B}(K)$ are isometric:

Theorem 32 (Riesz, Theorem 6.19, (Rud87)) *Let K be a compact Hausdorff space. Then, every bounded linear functional ϕ on $C(K)$ is represented by a unique regular finite Borel signed[9] mea-*

---

9. Theorem 6.19 in (Rud87) considers complex functionals and complex measures. However, one can show that if the functional is real then the corresponding regular measure must also be real. One way to see it is to use Urysohn's Lemma to show that a regular complex measure for which all continuous functions have real integrals is in fact a real measure.

sure μ, in the sense that

$$\phi f = \int_K f \, d\mu$$

for every $f \in C(K)$. Moreover, the operator norm of ϕ is the total variation of μ:

$$\|\phi\| = \|\mu\|_{TV}.$$

This natural identification between $C(K)^\star$ and $\mathcal{B}(K)$ allows us to define a topology on $\mathcal{B}(K)$ in terms of a topology on $C(K)^\star$. We do so by considering the weak* on $C(K)^\star$. Thus, this is the coarsest topology such that the operators

$$T_f(\mu) = \int f \, d\mu$$

are continuous for every $f \in C(K)$. We now use Banach-Alaoglu Theorem (Theorem 3.15 in (Rud91)) which implies that the closed unit ball of $C(K)^\star$ is closed with respect to the weak* topology. Consequently by the above identification of $C(K)^\star$ and $\mathcal{B}(K)$ we obtain:

Claim 33 *Let K be a compact Hausdorff space. Then the closed unit ball of $\mathcal{B}(K)$,*

$$B\big[\mathcal{B}(K)\big] := \{\mu \in \mathcal{B}(K) \mid \|\mu\|_{TV} \leq 1\},$$

is compact in the weak* topology.

C.2. Proof of Theorem 8 [Strong duality in the contradiction graph]

Theorem *[Theorem 8 Restatement] Let $\mathcal{X}$ be an arbitrary domain, $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ a concept class and $m \in \mathbb{N}$. Let $\omega_m^\star$ and $\chi_m^\star$ denote the fractional clique and chromatic numbers of the contradiction graph $G_m(\mathcal{H})$. Then,*

$$\omega_m^\star = \chi_m^\star \leq 2^m.$$

Moreover, there exists a fractional coloring realizing $\chi_m^\star$. (I.e. the infimum is in fact a minimum.)

Proof [Proof of Theorem 8] In order to prove the equality $\omega_m^\star = \chi_m^\star$ we use Sion's Theorem. Recall the extended definitions of fractional colorings and cliques of $G_m(\mathcal{H})$: fractional colorings are regular Borel probability measures on $\{0,1\}^X$ (where the latter is equipped with the product topology). The set of fractional colorings is denoted $\Delta(\{0,1\}^X)$. Fractional cliques are finitely supported distributions over $V_m(\mathcal{H})$, the set of $\mathcal{H}$-realizable datasets of size m. The set of fractional cliques is denoted $\Delta(V_m^{\mathcal{H}})$.

We plug in Sion's theorem $W = \Delta\big(\{0,1\}^{\mathcal{X}}\big)$, $U = \Delta(V_m^{\mathcal{H}})$, and

$$F(\mu, \nu) = \mathop{\mathbb{E}}_{\substack{h \sim \mu \\ S \sim \nu}} \left[\mathbb{1}[h \text{ is consistent with } S] \right].$$

Note that by definition of F,

$$\sup_{w \in W} \inf_{u \in U} F(w, u) = \frac{1}{\chi_m^\star},$$

$$\inf_{u \in U} \sup_{w \in W} F(w, u) = \frac{1}{\omega_m^\star}.$$

Applying Sion's Theorem with W, U and F as above proves the equality $\omega_m^\star = \chi_m^\star$, and the claim that there exists a fractional coloring realizing $\chi_m^\star$. In order to apply Sion's theorem on W, U, F,

we need to define the corresponding linear topological spaces and verify that the assumptions in the premise of Sion's theorem are satisfied.

Step 1: $\Delta\big(\{0,1\}^{\mathcal{X}}\big)$ **is a compact convex subset of a linear topological space.** $\Delta\big(\{0,1\}^{\mathcal{X}}\big)$ is a subset of $\mathcal{B}(\{0,1\}^{\mathcal{X}})$, which is the set of all signed finite regular Borel measures on $\{0,1\}^{\mathcal{X}}$. We consider $\mathcal{B}(\{0,1\}^{\mathcal{X}})$ as a linear topological space equipped with the weak* topology. Clearly, $\Delta(\{0,1\}^X)$ is convex, and thus it remains to show that it is compact. Notice that $\Delta(\{0,1\}^X)$ is contained in the unit ball $B[\mathcal{B}(\{0,1\}^X)]$, and the latter is compact by Claim 33. Thus, it remains to show that $\Delta(\{0,1\}^X)$ is closed in the weak* topology, which follows from the next claim.

Claim 34 *Let*

$$B = B\big[\mathcal{B}\big(\{0,1\}^{\mathcal{X}}\big)\big] = \Big\{\mu \in \mathcal{B}\big(\{0,1\}^{\mathcal{X}}\big) \mid \|\mu\|_{TV} \leq 1\Big\},$$

$$S = \Big\{\mu \in \mathcal{B}\big(\{0,1\}^X\big) \mid \mu\big(\{0,1\}^{\mathcal{X}}\big) = 1\Big\}.$$

Then, $\Delta\big(\{0,1\}^{\mathcal{X}}\big) = B \cap S$.

Note that the above claim implies that $\Delta\big(\{0,1\}^{\mathcal{X}}\big)$ is weak*-closed as an intersection of two closed subsets: B is weak*-compact (and in particular closed) by Claim 33. To see that S is closed, consider the operator $T_1(\mu) = \int 1 \, d\mu = \mu(\{0,1\}^X)$, which is weak*-continuous because the constant map 1 is continuous. Thus, $S = T_1^{-1}(\{1\})$ is closed.

Proof [Proof of Claim 34] It is clear that $\Delta\big(\{0,1\}^{\mathcal{X}}\big)$ is contained in the right-hand side. From the other direction, it is enough to show that if $\mu \in B \cap S$ then $\mu \geq 0$. Write $\mu = \mu^+ - \mu^-$, where μ^+, μ^- are the positive and negative variations of μ. Since $\mu \in B$,

$$\|\mu\|_{TV} = \|\mu^+\|_{TV} + \|\mu^-\|_{TV} \leq 1. \tag{22}$$

Since $\mu \in S$,

$$\begin{aligned}
1 = \mu(\{0,1\}^X) &= \mu^+(\{0,1\}^X) - \mu^-(\{0,1\}^X) \\
&= |\mu^+|(\{0,1\}^X) - |\mu^-|(\{0,1\}^X) \quad \text{(because } \mu^+, \mu^- \text{ are positive measures)} \\
&= \|\mu^+\|_{TV} - \|\mu^-\|_{TV}. \tag{23}
\end{aligned}$$

From Equations (22) and (23) we conclude $\|\mu^-\|_{TV} = 0$, i.e. $\mu^- = 0$, and $\mu = \mu^+$ is a positive, completing the proof. ∎

Step 2: $\Delta(V_m^{\mathcal{H}})$ **is a convex subset of a linear topological space.** Consider the space of all finitely supported signed measures on $V_m(\mathcal{H})$ with the discrete topology. $\Delta(V_m^{\mathcal{H}})$ is a convex subset of this space.

Step 3: F **is linear and continuous in each coordinate.** First, notice that $F(\mu, \nu)$ is linear in each coordinate by linearity of expectation with respect to the underlying measure.

We next show that $F(\cdot, \nu)$ is continuous for every fixed ν. By Remark 29, for every fixed ν, $F(\cdot, \nu)$ is a finite convex combination of functions of the form

$$F_S(\mu) = \mathop{\mathbb{E}}_{h \sim \mu}\big[\mathbb{1}[h \text{ is consistent with } S]\big],$$

where S is a fixed $\mathcal{H}$-realizable dataset of size m. Thus it suffices to show that $F_S(\mu)$ is weak*-continuous for each fixed S. Let $f_S(h) = \mathbb{1}[h \text{ is consistent with } S]$ and notice that $F_S(\mu) =$

$\mathbb{E}_{h \sim \mu}[f_S(h)]$. Thus, by the definition of the weak* topology, it is enough to show that the map f_S is continuous with respect to the product topology on $\{0,1\}^{\mathcal{X}}$. Since f_S is an indicator map (with values in $\{0,1\}$), continuity amounts to showing that both $f_S^{-1}(0)$ and $f_S^{-1}(1)$ are open sets. Indeed, by definition $f_S^{-1}(1)$ is a basic open set which is also closed and hence $f_S^{-1}(0)$ is also open as required. It remains to show that $F(\mu, \cdot)$ is continuous in its second coordinate for every fixed μ. This is straightforward because the topology on the second coordinate $(\Delta(V_m^{\mathcal{H}}))$ is discrete and hence every function is continuous.

To complete the proof of Theorem 8, it is left to show that $\chi^m \leq 2^m$. Recall,

$$\frac{1}{\chi_m^\star} = \sup_{\mu \in \Delta\left(\{0,1\}^{\mathcal{X}}\right)} \inf_{\nu \in \Delta(V_m^{\mathcal{H}})} \mathbb{E}_{\substack{h \sim \mu, \\ S \sim \nu}} \left[\mathbb{1}[h \text{ is consistent with } S] \right].$$

Hence, it suffices to show that there exists a distribution $\mu^\star \in \Delta\left(\{0,1\}^{\mathcal{X}}\right)$ such that for every realizable dataset $S = \left((x_1, y_1), \ldots, (x_m, y_m)\right) \in V_m(\mathcal{H})$,

$$\Pr_{h \sim \mu^\star}[h \text{ is consistent with } S] = \mu^\star\left(\{h \mid h(x_i) = y_i, i = 1, \ldots, m\}\right) = \frac{1}{2^m}.$$

Indeed if this holds, then by definition $\frac{1}{\chi_m^\star} \geq \frac{1}{2^m}$. Therefore, the following lemma concludes the proof.

Lemma 35 *Let $\mathcal{X}$ be an arbitrary domain. Then there exists a regular Borel probability measure $\mu^\star$ over $\{0,1\}^{\mathcal{X}}$, which satisfies the following property. For every finite set $\{x_1, \ldots, x_k\} \subseteq \mathcal{X}$ and labels $y_1, \ldots, y_k$*

$$\mu^\star\left(\{h \in \{0,1\}^{\mathcal{X}} \mid h(x_i) = y_i, i = 1, \ldots, k\}\right) = \frac{1}{2^k}.$$

Before turning to prove Lemma 35, we first need to recollect some definitions and state known results from measure theory which are needed for the proof. The proof is deferred to Section C.3. ∎

C.3. Tossing a fair coin infinitely many times

One of the basic facts we used in the paper is that the fractional chromatic number of the contradiction graph is bounded. To prove this fact we drew a random hypothesis by sampling a value independently for each $x \in \mathcal{X}$ uniformly over $\{0,1\}$.

For a countable domain $\mathcal{X}$ it is clear that this sampling process induces a well-defined probability distribution over $\{0,1\}^{\mathcal{X}}$. However, if $\mathcal{X}$ is arbitrary, and possibly uncountable, it is not clear that this process is even well-defined.

In this section, we prove Lemma 35: we show that indeed the described sampling process induces a probability measure on the product space $\{0,1\}^{\mathcal{X}}$. Furthermore, we show that this measure is a regular Borel measure, and hence a valid fractional coloring. (Recall that we identify fractional colorings as regular Borel probability measures over $\{0,1\}^{\mathcal{X}}$, where the latter is equipped with the product topology).

C.3.1. Measure Theory Preliminaries

The Product σ-Algebra. Let $(X_\alpha, \Sigma_\alpha)_{\alpha \in \mathcal{I}}$ be a family of measurable spaces. Denote by $X_{\mathcal{I}} = \prod_{\alpha \in \mathcal{I}} X_\alpha$ the product space, and by $\pi_\alpha : X_{\mathcal{I}} \to X_\alpha$ the projection map to X_α. Using the projection

π_α, we can pull back the σ-algebra Σ_α on X_α, to a σ-algebra on $X_\mathcal{I}$:

$$\pi_\alpha^*(\Sigma_\alpha) := \{\pi_\alpha^{-1}(E_\alpha) \mid E_\alpha \in \Sigma_\alpha\}.$$

Define the *product σ-algebra* on $X_\mathcal{I}$, denoted $\Sigma_\mathcal{I}$, to be the σ-algebra generated by all $\pi_\alpha^*(\Sigma_\alpha)$:

$$\Sigma_\mathcal{I} := \left\langle \bigcup_{\alpha \in \mathcal{I}} \pi_\alpha^*(\Sigma_\alpha) \right\rangle = \langle \pi_\alpha^{-1}(E_\alpha) \mid \alpha \in \mathcal{I} \rangle.$$

Using similar notations, for every subset of indices $A \subseteq \mathcal{I}$, let $X_A = \prod_{\alpha \in A} X_\alpha$, let $\pi_A : X_\mathcal{I} \to X_A$ denote the projection to X_A, and denote the product σ-algebra on A by

$$\Sigma_A = \left\langle \bigcup_{\alpha \in A} \pi_A \circ \pi_\alpha^*(\Sigma_\alpha) \right\rangle.$$

Given $A \subseteq \mathcal{I}$, a measure $\mu_\mathcal{I}$ on the product σ-algebra $\Sigma_\mathcal{I}$, induces a measure μ_A on Σ_A (which is the pushforward of $\mu_\mathcal{I}$):

$$\mu_A(E) := (\pi_A)_* \mu_\mathcal{I}(E) = \mu_\mathcal{I}(\pi_A^{-1}(E)).$$

Note that those measures obey the compatibility relation: For $B \subseteq A \subseteq I$, denote by $\pi_{A \to B}$ the projection $\pi_B|_{X_A} : X_A \to X_B$. Then

$$\mu_B = (\pi_{A \to B})_* \mu_A, \tag{24}$$

i.e, for every $E \in \Sigma_B$, $\mu_B(E) = \mu_A((\pi_{A \to B})^{-1}(E))$.

Kolmogorov's Extension Theorem.   A natural question is whether one can reconstruct $\mu_\mathcal{I}$ solely from the projections μ_A to *finite* subsets $A \subseteq \mathcal{I}$. It turns out that in the special case where the μ_A are probability measures (and satisfy some additional regularity and compatibility conditions), it is indeed possible. This is the content of Kolmogorov's Theorem (Theorem 2.4.3 in (Tao11)).

Theorem 36 (Kolmogorov's Extension Theorem)   *Let $(X_\alpha, \Sigma_\alpha, \tau_\alpha)_{\alpha \in \mathcal{I}}$ be a family of measurable spaces $(X_\alpha, \Sigma_\alpha)$, equipped with a topology τ_α. For every finite subset $A \subseteq \mathcal{I}$, let μ_A be a probability measure on Σ_A which is inner regular with respect to the product topology $\tau_A := \prod_{\alpha \in A} \tau_\alpha$ on X_A. Furthermore, assume μ_A, μ_B obey the compatibility condition in Equation (24) for every nested finite subsets $B \subseteq A$ of $\mathcal{I}$. Then, there exists a unique probability measure $\mu_\mathcal{I}$ on the product σ-algebra $\Sigma_\mathcal{I}$, with the property that $(\pi_A)_* \mu_\mathcal{I} = \mu_A$ for all finite subset $A \subseteq \mathcal{I}$. (I.e. the push-forward measure of $\mu_\mathcal{I}$ when projected on A equals to μ_A.)*

Extending a product measure to a regular Borel measure.   Let us assume here and below that each X_α is a compact Hausdorff space, and Σ_α is the Borel σ-algebra on X_α. Kolmogorov's theorem allows to construct a probability measure on the product σ-algebra by specifying its behavior over finite projections. The following theorem allows us to further extend this measure to the Borel σ-algebra of the product space $X_\mathcal{I}$. (Notice that the Borel σ-algebra is finer than the Product σ-algebra.)

Theorem 37 (Theorem 2, (Kak43))   *Let $(X_\alpha, \Sigma_\alpha)_{\alpha \in \mathcal{I}}$ be a family measurable spaces where each X_α is a compact Hausdorff space and Σ_α is the Borel σ-algebra of X_α. Then, every probability measure on the product σ-algebra, $\Sigma_\mathcal{I}$, has a unique extension to a regular Borel measure $\mu_\mathcal{I}^*$ on $X_\mathcal{I}$.*

C.3.2. PROOF OF LEMMA 35

Proof For every finite set of unlabeled examples $\{x_1, \ldots, x_k\} \subseteq \mathcal{X}$, let $\mu_{\{x_1, \ldots, x_k\}}$ be the uniform distribution measure over $\{0,1\}^{\{x_1, \ldots, x_k\}}$. Notice that $\{\mu_{\{x_1, \ldots, x_k\}} : \{x_1, \ldots, x_k\} \subseteq \mathcal{X}\}$ are regular and obey the the compatibility condition in Equation (24). By Theorem 36 there exists a probability measure μ over the product σ-algebra of $\{0,1\}^{\mathcal{X}}$ which satisfies the equations stated in the Lemma. By Theorem 37 there exists a regular Borel probability measure $\mu^\star$ as required. ∎

Appendix D. Additional Proofs

D.1. Proof of Lemma 1 [Clique and fractional clique numbers are bounded]

Lemma *[Lemma 1, Restatement] Let ω_m and $\omega_m^\star$ denote the clique and fractional clique numbers of $G_m(\mathcal{H})$. Then, $\omega_m \leq \omega_m^\star \leq 2^m$.*

We note that Lemma 1 holds in a more general setting as well, where the domain $\mathcal{X}$ is arbitrary (possibly uncountable). We refer the reader to Appendix C for more details.

Proof Let δ be a fractional clique in $G_m(\mathcal{H})$. Draw a random $h \in \{0,1\}^{\mathcal{X}}$ such that $h(x)$ is drawn uniformly and independently from $\{0,1\}$ for each $x \in \mathcal{X}$. Consider the random variable

$$X = \sum_{S \in V_m(\mathcal{H})} \mathbb{1}\left[S \text{ is consistent with } h\right] \cdot \delta(S).$$

Note that $X \leq 1$ almost surely: indeed, the set $\{S : S \text{ is consistent with } h\}$ is independent for every hypothesis h. Therefore,

$$
\begin{aligned}
1 \geq \mathbb{E}\left[X\right] &= \sum_{S \in V_m(\mathcal{H})} \Pr\left[S \text{ is consistent with } h\right] \cdot \delta(S) \\
&= \sum_{S \in V_m(\mathcal{H})} 2^{-m} \cdot \delta(S) \\
&= 2^{-m} |\delta|.
\end{aligned}
$$

Thus $|\delta| \leq 2^m$ and therefore $\omega_m^\star \leq 2^m$. Note that the inequality $\omega_m \leq \omega_m^\star$ trivially holds since every (integral) clique is a fractional clique. ∎

D.2. Proof of Lemma 11 [Independent sets and consistent hypotheses]

Lemma *[Lemma 11, Restatement] Let $\mathcal{H}$ be a class and m be a natural number.*

1. *For every independent set I in $G_m(\mathcal{H})$, there exists an hypothesis $h \in \{0,1\}^{\mathcal{X}}$ such that h is consistent with every dataset $S \in I$; i.e. for every dataset $S \in I$ and every example $(x,y) \in S$, we have $h(x) = y$.*

2. *For every hypothesis h, the set of all datasets of size m that are consistent with h is an independent set in $G_m(\mathcal{H})$; i.e. the set*

$$V_h := \{S = ((x_1, y_1), \ldots, (x_m, y_m)) \in V_m(\mathcal{H}) \mid \forall i, h(x_i) = y_i\}$$

is independent in $G_m(\mathcal{H})$

Proof First, we prove the first part of the lemma. Let I be an independent set in $G_m(\mathcal{H})$ and denote
$$L(I) = \big\{(x,y) \in \mathcal{X} \times \{0,1\} \mid \exists S \in I \text{ s.t. } (x,y) \in S\big\}.$$
Observe that since I is independent, if $(x,y) \in L(I)$ then $(x, 1-y)$ must not be in $L(I)$ (because datasets containing $(x, 1-y)$ are connected with an edge to datasets containing (x,y)). Define an hypothesis h as follows: for every $(x,y) \in L(I)$ set $h(x) = y$. For every x such that neither $(x,0)$ or $(x,1)$ are in $L(I)$, set arbitrarily $h(x) = 0$. Indeed h interpolates every dataset $S \in I$.

Next, we prove the second part of the lemma. Let $h \in \{0,1\}^{\mathcal{X}}$ be a hypothesis. V_h is indeed independent in $G_m(\mathcal{H})$, else there must be $S, S' \in V_h$ and an unlabeled example $x \in \mathcal{X}$ such that $(x,0) \in S, (x,1) \in S'$, and by definition $h(x) = 0 = 1$, leading to a contradiction. ∎

D.3. Proof of Lemma 12 [Fractional cliques and colorings vs. distributions]

Lemma *[Lemma 12, Restatement] Let $\mathcal{H}$ be a class, $m \in \mathbb{N}$. Then,*

1. *There exists a fractional coloring c of $G_m(\mathcal{H})$ with $\mathtt{col}(c) = \alpha > 0$ if and only if there exists a distribution μ over hypotheses such that*
$$\inf_{S} \Pr_{h \sim \mu} [h \text{ is consistent with } S] = \frac{1}{\alpha},$$
   where the infimum is taken over realizable datasets of size m.

2. *There exists a fractional clique δ of $G_m(\mathcal{H})$ with $|\delta| = \alpha > 0$ if and only if there exists a distribution ν over realizable datasets of size m such that*
$$\sup_{h} \Pr_{S \sim \nu} [h \text{ is consistent with } S] = \frac{1}{\alpha},$$
   where the supremum is taken over hypotheses $h \in \{0,1\}^{\mathcal{X}}$.

Proof As demonstrated in Section 3.2, there is a correspondence between fractional colorings and independent sets: a fractional coloring of $G_m(\mathcal{H})$ with α colors corresponds to a distribution μ over independent sets of $G_m(\mathcal{H})$ with
$$\mathtt{val}(\mu) = \frac{1}{\alpha} = \inf_{S \in V_m(\mathcal{H})} \Pr_{I \sim \mu} [S \in I].$$
From Lemma 11, there is a correspondence between (maximal) independent sets and hypotheses: an independent set I in $G_m(\mathcal{H})$ corresponds to a hypothesis h which is consistent with all datasets $S \in I$. Note that a realizable dataset S of size m belongs to an independent set I if and only if S is consistent with the hypothesis h corresponds to I. Hence, by abuse of notation, consider μ to be a distribution over hypotheses. Observe that for every realizable dataset S of size m,
$$\Pr_{h \sim \mu} [h \text{ is consistent with } S] = \Pr_{I \sim \mu} [S \in I].$$
This completes Item 1 in the lemma. Note that Item 2 is immediate, again it follows from the discussion in Section 3.2, since the vertices of the contradiction are realizable datasets of size m. ∎

D.4. Proof of Lemma 20

Lemma *[Lemma 20, Restatement] Let $\mathcal{H}$ be a hypothesis class. Then*
$$\mathtt{CD}(\mathcal{H}) \leq \max\big\{2\mathtt{LD}(\mathcal{H}) \log(\mathtt{LD}(\mathcal{H})), 300\big\}.$$

To show the desired bound we will use the following technical lemma.

Lemma 38 *Let $d \in \mathbb{N}$ and $m_0 \in \mathbb{R}$ such that*

(i) $m_0 \geq \frac{d}{\ln 2}$, and
(ii) $2^{m_0} \geq (2m_0 + 1)^d$.

Then,

$$(\forall m > m_0) : 2^m > (2m + 1)^d.$$

Further, $m_0 = 2d \log d$ satisfies items (i) and (ii) above, provided that $d \geq 30$.

Proof [Proof of Lemma 20] If $\mathrm{LD}(\mathcal{H}) = \infty$ then the inequality trivially holds. Suppose $\mathrm{LD}(\mathcal{H}) = d < \infty$. We distinguish between two cases: if $d < 30$ then by Lemma 18,

$$\omega_m \leq (2m + 1)^d < (2m + 1)^{30}$$

which is less than 2^m for every $m \geq 300$, and thus $\mathrm{CD}(\mathcal{H}) \leq 300 \leq \max\{2d \log d, 300\}$ as required. Else, if $d \geq 30$ then by Lemma 38 for every $m > 2d \log d$

$$\omega_m \leq (2m + 1)^d < 2^m,$$

thus $\mathrm{CD}(\mathcal{H}) \leq 2d \log(d) \leq \max\{2d \log d, 300\}$ as required. ∎

Proof [Proof of Lemma 38] We first show that $2^m > (2m+1)^d$ for all $m > m_0$. Indeed, if $m > m_0$ then

$$2^m = 2^{m-m_0} \cdot 2^{m_0}$$
$$\geq 2^{m-m_0} \cdot (2m_0 + 1)^d. \qquad \text{(by assumption (ii))}$$

Note that

$$\left(\frac{2m + 1}{2m_0 + 1}\right)^d = \left(1 + \frac{2(m - m_0)}{2m_0 + 1}\right)^d$$
$$< \exp\left(\frac{2(m - m_0) \cdot d}{2m_0 + 1}\right) \qquad (1 + x < \exp(x) \text{ for all } x > 0)$$
$$\leq \exp\left(\ln(2)(m - m_0)\right) \qquad (\tfrac{2d}{2m_0+1} \leq \ln 2 \text{ by assumption (i)})$$
$$= 2^{m-m_0},$$

which implies that $2^{m-m_0} \cdot (2m_0 + 1)^d > (2m + 1)^d$, therefore by the previous derivation, $2^m > (2m+1)^d$ as wanted. Next, assume $d \geq 30$ and set $m_0 = 2d \log d$. We will show m_0 satisfies items (i) and (ii). We first prove that $m_0 \geq \frac{d}{\ln 2}$:

$$m_0 \geq \frac{d}{\ln 2} \iff 2d \log d \geq \frac{d}{\ln 2}$$
$$\iff \log d \geq \frac{1}{2 \ln 2}$$
$$\iff d \geq 2^{\frac{1}{2\ln 2}},$$

and indeed $d \geq 2^{\frac{1}{2\ln 2}} = \sqrt{e}$ since $d \geq 30$. Lastly, we prove that $2^{m_0} \geq (2m_0 + 1)^d$:

$$
\begin{aligned}
2^{m_0} = 2^{2d \log d} &= d^{2d} \\
&\geq (6d \log d)^d &&\text{(since } d \geq 6 \log d \text{ for all } d \geq 30) \\
&= (3m_0)^d \\
&\geq (2m_0 + 1)^d.
\end{aligned}
$$

∎

D.5. Proof of Lemma 23

Lemma *[Lemma 23, Restatement] Let $\alpha \geq 2$, and set $m = \lfloor 20\alpha \ln \alpha \rfloor$ and $k = 4m^\alpha$. Then,*

$$
\left(1 - q(m)\right)^k \leq \frac{1}{4},
$$

where $q(m) = \frac{1}{m^\alpha} - \left(\frac{3}{4}\right)^m$.

Proof Observe that

$$
(1 - q(m))^k \leq \exp(-k \cdot q(m)) \qquad (1 + x < \exp(x) \text{ for all } x > 0)
$$

Hence it is enough to show that for $m = \lfloor 20\alpha \ln \alpha \rfloor$, $k = 4m^\alpha$

$$
k \cdot q(m) \geq \ln 4.
$$

Indeed, it easy to verify that $\left(\frac{3}{4}\right)^m \leq \frac{1}{2m^\alpha}$ where $\alpha \geq \alpha$ and $m = \lfloor 20\alpha \ln \alpha \rfloor$. Therefore,

$$
\begin{aligned}
k \cdot q(m) = k \left(\frac{1}{m^\alpha} - \left(\frac{3}{4}\right)^m \right) &\geq k \cdot \frac{1}{2m^\alpha} \\
&= 2 &&\text{(setting } k = 4m^\alpha) \\
&\geq \ln 4.
\end{aligned}
$$

∎