In *Proc. of 20th Conference on Learning Theory (COLT),* June 2007.

A Lower Bound for Agnostically Learning Disjunctions

Adam R. Klivans Alexander A. Sherstov

The University of Texas at Austin
Department of Computer Sciences
Austin, TX 78712 USA
{klivans,sherstov}@cs.utexas.edu

Abstract. We prove that the concept class of disjunctions cannot be pointwise approximated by linear combinations of any small set of *arbitrary* real-valued functions. That is, suppose there exist functions $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ with the property that every disjunction f on n variables has $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant 1/3$ for some reals $\alpha_1, \ldots, \alpha_r$. We prove that then $r \geqslant 2^{\Omega(\sqrt{n})}$. This lower bound is tight. We prove an incomparable lower bound for the concept class of linear-size DNF formulas. For the concept class of majority functions, we obtain a lower bound of $\Omega(2^n/n)$, which almost meets the trivial upper bound of 2^n for *any* concept class.

These lower bounds substantially strengthen and generalize the polynomial approximation lower bounds of Paturi and show that the regression-based agnostic learning algorithm of Kalai et al. is optimal. Our techniques involve a careful application of results in communication complexity due to Razborov and Buhrman et al.

1 Introduction

Approximating Boolean functions by linear combinations of small sets of features is a fundamental area of study in machine learning. Well-known algorithms such as linear regression, support vector machines, and boosting attempt to learn concepts as linear functions or thresholds over a fixed set of real-valued features.

In particular, much work in learning theory has centered around approximating various concept classes, with respect to a variety of distributions and metrics, by *low-degree polynomials* [3,10,17–19,21,26,28]. In this case, the features mentioned above are simply monomials. For example, Linial et al. [21] gave a celebrated uniform-distribution algorithm for learning constant-depth circuits by proving that any such circuit can be approximated by a low-degree Fourier polynomial, with respect to the uniform distribution and ℓ_2 norm.

A more recent application of this polynomial technique is due to Kalai et al. [11], who considered the well-studied problem of agnostically learning disjunctions [6, 14, 27, 34]. Kalai et al. recalled a result of Paturi [29] that a disjunction on n variables can be approximated pointwise by a degree-$\tilde{O}(\sqrt{n})$ polynomial. They then used linear regression to obtain the first subexponential ($2^{\tilde{O}(\sqrt{n})}$-time) algorithm for *agnostically* learning disjunctions with respect to *any* distribution [11, Thm. 2]. More generally, Kalai et al. used ℓ_∞-norm approximation to formulate the first, and so far only, approach

to distribution-free agnostic learning. One goal of this paper is to show the fundamental limits of this approximation-based paradigm.

1.1 Key Definitions

Before stating our results formally, we briefly describe our notation. A Boolean function is a mapping $f : \{-1,1\}^n \to \{-1,1\}$, where -1 corresponds to "true." A *feature* is any function $\phi : \{-1,1\}^n \to \mathbb{R}$. We say that ϕ *approximates* f *pointwise within* ε, denoted

$$\|f - \phi\|_\infty \leqslant \varepsilon,$$

if $|f(x) - \phi(x)| \leqslant \varepsilon$ for all x. We say that *a linear combination of features* $\phi_1, \ldots, \phi_r$ *approximates* f *pointwise within* ε if $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant \varepsilon$ for some reals $\alpha_1, \ldots, \alpha_r$.

1.2 Our Results

Let $\mathscr{C}$ be a concept class. Suppose that $\phi_1, \ldots, \phi_r$ are features whose linear combinations can pointwise approximate every function in $\mathscr{C}$. We first observe that the algorithm of Kalai et al.—assuming that $\phi_1, \ldots, \phi_r$ can be evaluated efficiently—learns $\mathscr{C}$ agnostically under any distribution in time $\mathsf{poly}(n, r)$. As far as we are aware, this is the only known method for developing provably efficient, distribution-free agnostic learning algorithms. To determine the limits of this paradigm, our paper focuses on lower bounds on r for an *arbitrary* choice of features.

We start with the concept class of disjunctions.

Theorem 1 (Disjunctions). *Let $\mathscr{C} = \{\bigvee_{i \in S} x_i : S \subseteq [n]\}$ be the concept class of disjunctions. Let $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ be arbitrary functions whose linear combinations can pointwise approximate every $f \in \mathscr{C}$ within $\varepsilon = 1/3$. Then $r \geqslant 2^{\Omega(\sqrt{n})}$.*

Theorem 1 obviously also holds for the concept class of *conjunctions*.

Theorem 1 shows the optimality of using monomials as features for approximating disjunctions. In particular, it rules out the possibility of using the algorithm of Kalai et al. with other, cleverly constructed features to obtain an improved agnostic learning result for disjunctions.

We obtain an incomparable result against linear-size DNF formulas.

Theorem 2 (DNF formulas). *Let $\mathscr{C}$ be the concept class of DNF formulas of linear size. Let $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ be arbitrary functions whose linear combinations can pointwise approximate every $f \in \mathscr{C}$ within $\varepsilon = 1 - 2^{-cn^{1/3}}$, where $c > 0$ is a sufficiently small absolute constant. Then $r \geqslant 2^{\Omega(n^{1/3})}$.*

Theorems 1 and 2 both give exponential lower bounds on r. Comparing the two, we see that Theorem 1 gives a better bound on r against a simpler concept class. On the other hand, Theorem 2 remains valid for a particularly weak success criterion: when the approximation quality is exponentially close to trivial ($\varepsilon = 1$).

The last concept class we study is that of majority functions. Here we prove our best lower bound, $r = \Omega(2^n/n)$, that essentially meets the trivial upper bound of 2^n for any

concept class. Put differently, we show that the concept class of majorities is essentially as hard to approximate as *any* concept class at all. In particular, this shows that the Kalai et al. paradigm cannot yield any nontrivial ($2^{o(n)}$-time) distribution-free algorithm for agnostically learning majority functions.

Theorem 3 (Majority functions). *Let $\mathscr{C} = \{\mathrm{MAJ}(\pm x_1, \ldots, \pm x_n)\}$ be the concept class of majority functions. Let $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ be arbitrary functions whose linear combinations can pointwise approximate every $f \in \mathscr{C}$ within $\varepsilon = c/\sqrt{n}$, where c is a sufficiently small absolute constant. Then $r \geqslant \Omega(2^n/n)$. For approximation to within $\varepsilon = 1/3$, we obtain $r \geqslant 2^{\Omega(n/\log n)}$.*

We also relate our inapproximability results to the fundamental notions of *dimension complexity* and *SQ dimension* (Sections 5–7). Among other things, we show that the types of approximation lower bounds we study are prerequisites for lower bounds on dimension complexity and the SQ dimension. It is a hard open problem [32] to prove exponential lower bounds on the dimension complexity and SQ dimension of polynomial-size DNF formulas, or even AC^0 circuits.

Optimality of polynomial-based approximation. The preceding discussion has emphasized the implications of Theorems 1–3 in learning theory. Our results also have interesting consequences in approximation theory. Paturi [29] constructs polynomials of degree $\tilde{\Theta}(\sqrt{n})$ and $\Theta(n)$ that pointwise approximate disjunctions and majority functions, respectively. He also shows that these *degree* results are optimal for polynomials. This, of course, does not exclude polynomials that are *sparse*, i.e., contain few monomials. Our lower bounds strengthen Paturi's result by showing that the approximating polynomials cannot be sparse. In addition, our analysis remains valid when monomials are replaced by *arbitrary* features. As anticipated, our techniques differ significantly from Paturi's.

1.3 Our Techniques

To prove our approximation lower bounds, we need to use various techniques from matrix analysis, communication complexity, and Fourier analysis. We obtain our main theorems in two steps. First, we show how to place a lower bound on the quantity of interest (the size of feature sets that pointwise approximate a concept class $\mathscr{C}$) using the *discrepancy* and the *ε-approximate trace norm* of the characteristic matrix of $\mathscr{C}$. The latter two quantities have been extensively studied. In particular, the discrepancy estimate that we need is a recent result of Buhrman et al. [5]. For estimates of the ε-approximate trace norm, we turn to the pioneering work of Razborov [30] on quantum communication complexity, as well as a recent construction of Linial and Shraibman [24].

2 Preliminaries

The notation $[n]$ stands for the set $\{1,2,\ldots,n\}$, and $\binom{[n]}{k}$ stands for the family of all k-element subsets of $[n] = \{1,2,\ldots,n\}$. The symbol $\mathbb{R}^{n\times m}$ refers to the family of all $m \times n$ matrices with real entries. The (i,j)th entry of a matrix A is denoted by A_{ij} or

$A(i,j)$. We frequently use "generic-entry" notation to specify a matrix succinctly: we write $A = [F(i,j)]_{i,j}$ to mean that that the (i,j)th entry of A is given by the expression $F(i,j)$.

A *concept class* $\mathscr{C}$ is any set of Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$. The *characteristic matrix* of $\mathscr{C}$ is the matrix $M = [f(x)]_{f\in\mathscr{C},\, x\in\{-1,1\}^n}$. In words, the rows of M are indexed by functions $f \in \mathscr{C}$, the columns are indexed by inputs $x \in \{-1,1\}^n$, and the entries are given by $M_{f,x} = f(x)$.

2.1 Agnostic Learning

The agnostic learning model was defined by Kearns et al. [15]. It gives the learner access to arbitrary example-label pairs with the requirement that the learner output a hypothesis competitive with the best hypothesis from some fixed concept class. Specifically, let D be a distribution on $\{-1,1\}^n \times \{-1,1\}$ and let $\mathscr{C}$ be a concept class. For a Boolean function f, define its *error* as $\mathrm{err}(f) = \Pr_{(x,y)\sim D}[f(x) \neq y]$. Define the *optimal error* of $\mathscr{C}$ as $\mathsf{opt} = \min_{f\in\mathscr{C}} \mathrm{err}(f)$.

A concept class $\mathscr{C}$ is *agnostically learnable* if there exists an algorithm which takes as input δ, ε, and access to an example oracle $\mathsf{EX}(D)$, and outputs with probability at least $1-\delta$ a hypothesis $h : \{-1,1\}^n \to \{-1,1\}$ such that $\mathrm{err}(h) \leqslant \mathsf{opt} + \varepsilon$. We say $\mathscr{C}$ is agnostically learnable in time t if its running time (including calls to the example oracle) is bounded by $t(\varepsilon, \delta, n)$.

The following proposition relates pointwise approximation by linear combinations of features to efficient agnostic learning.

Proposition 1. *Fix $\varepsilon > 0$ and a concept class $\mathscr{C}$. Assume there are functions $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ whose linear combinations can pointwise approximate every $f \in \mathscr{C}$. Assume further that each $\phi_i(x)$ is computable in polynomial time. Then $\mathscr{C}$ is agnostically learnable to accuracy ε in time $\mathsf{poly}(r,n)$.*

We defer a proof of Proposition 1 to the full version. The needed simulation is a straightforward generalization of the ℓ_1 polynomial regression algorithm from Kalai et al. [11].

2.2 Fourier Transform

Consider the vector space of functions $\{-1,1\}^n \to \mathbb{R}$, equipped with the inner product $\langle f, g\rangle = 2^{-n}\sum_{x\in\{-1,1\}^n} f(x)g(x)$. The parity functions $\chi_S(x) = \prod_{i\in S} x_i$, where $S \subseteq [n]$, form an orthonormal basis for this inner product space. As a result, every Boolean function f can be uniquely written as

$$f = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S,$$

where $\hat{f}(S) = \langle f, \chi_S\rangle$. The f-specific reals $\hat{f}(S)$ are called the *Fourier coefficients* of f. We denote

$$\|\hat{f}\|_1 = \sum_{S\subseteq[n]} |\hat{f}(S)|.$$

2.3 Matrix Analysis

We draw freely on basic notions from matrix analysis; a standard reference on the subject is [9]. This section only reviews the notation and the more substantial results.

Let $A \in \mathbb{R}^{m\times n}$. We let $\|A\|_\infty \stackrel{\text{def}}{=} \max_{ij}|A_{ij}|$, the largest absolute value of an entry of A. We denote the singular values of A by $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \ldots \geqslant \sigma_{\min\{m,n\}}(A) \geqslant 0$. Recall that $\|A\|_\Sigma = \sum_{i=1}^{\min\{m,n\}} \sigma_i(A)$ and $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n A_{ij}^2}$ are the trace norm and Frobenius norm of A. We will also need the *ε-approximate trace norm,* defined as

$$\|A\|_\Sigma^\varepsilon = \min\{\|B\|_\Sigma : \|A-B\|_\infty \leqslant \varepsilon\}.$$

The well-known Hoffman-Wielandt inequality plays an important role in our analysis. In words, it states that small perturbations to the entries of a matrix result in small perturbations to its singular values. This inequality has seen numerous uses in the literature [8, 12, 25].

Theorem 4 (Hoffman-Wielandt inequality [9, Thm. 8.6.4]). *Let $A, B \in \mathbb{R}^{m\times n}$. Then $\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A) - \sigma_i(B))^2 \leqslant \|A-B\|_F^2$. In particular, if $\operatorname{rank}(B) = k$ then $\sum_{i\geqslant k+1} \sigma_i(A)^2 \leqslant \|A-B\|_F^2$.*

The Hoffman-Wielandt inequality is central to the following lemma, which allows us to easily construct matrices with high $\|\cdot\|_\Sigma^\varepsilon$ norm.

Lemma 1 (Linial and Shraibman [24], implicit). *Let $M = [f(x\oplus y)]_{x,y}$, where $f : \{-1,1\}^n \to \{-1,1\}$ is arbitrary. Then for all $\varepsilon \geqslant 0$,*

$$\|M\|_\Sigma^\varepsilon \geqslant 2^n(\|\hat{f}\|_1 - \varepsilon 2^{n/2}).$$

Proof (adapted from Linial and Shraibman [24]). Let $N = 2^n$ be the order of M. Consider an arbitrary matrix A with $\|A-M\|_\infty \leqslant \varepsilon$. We have:

$$N^2\varepsilon^2 \;\geqslant\; \|A-M\|_F^2 \quad \stackrel{\text{Thm. 4}}{\geqslant} \quad \sum_{i=1}^N (\sigma_i(A) - \sigma_i(M))^2 \;\geqslant\; \frac{1}{N}(\|A\|_\Sigma - \|M\|_\Sigma)^2,$$

so that $\|A\|_\Sigma \geqslant \|M\|_\Sigma - N^{3/2}\varepsilon$. Since the choice of A was arbitrary, we conclude that

$$\|M\|_\Sigma^\varepsilon \geqslant \|M\|_\Sigma - N^{3/2}\varepsilon. \tag{1}$$

It remains to analyze $\|M\|_\Sigma$. Let $Q = N^{-1/2}[\chi_S(x)]_{x,S}$. It is easy to check that Q is orthogonal. On the other hand,

$$M = [f(x\oplus y)]_{x,y} = \left[\sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x)\chi_S(y)\right]_{x,y} = Q \begin{bmatrix} N\hat{f}(\emptyset) & & \\ & \ddots & \\ & & N\hat{f}([n]) \end{bmatrix} Q^{\mathsf{T}}.$$

The last equation reveals the singular values of M. In particular, $\|M\|_\Sigma = N\|\hat{f}\|_1$. Together with (1), this completes the proof. □

A *sign matrix* is any matrix with ± 1 entries.

2.4 Communication Complexity

We consider functions $f : X \times Y \to \{-1,1\}$. Typically $X = Y = \{-1,1\}^n$, but we also allow X and Y to be arbitrary sets, possibly of unequal cardinality. A *rectangle* of $X \times Y$ is any set $R = A \times B$ with $A \subseteq X$ and $B \subseteq Y$. For a fixed distribution μ over $X \times Y$, the *discrepancy* of f is defined as

$$\mathrm{disc}_\mu(f) = \max_R \left| \sum_{(x,y)\in R} \mu(x,y) f(x,y) \right|,$$

where the maximum is taken over all rectangles R. We define $\mathrm{disc}(f) = \min_\mu \{\mathrm{disc}_\mu(f)\}$. We identify the function f with its *communication matrix* $M = [f(x,y)]_{x,y}$ and define $\mathrm{disc}_\mu(M) = \mathrm{disc}_\mu(f)$.

Discrepancy is a powerful quantity with various applications. In particular, it immediately yields lower bounds in various models of communication complexity, as well as circuit lower bounds for depth-2 majority circuits [20, 24, 33]. This paper shows yet another application of discrepancy. A definitive resource for further details on communication complexity is the book of Kushilevitz and Nisan [20].

2.5 SQ Dimension

The statistical query (SQ) model of learning, due to Kearns [13], is a restriction of Valiant's PAC model. See [16] for a comprehensive treatment. The SQ model is recognized as a powerful abstraction of learning and plays a major role in learning theory. The *SQ dimension* of $\mathscr{C}$ under μ, denoted $\mathrm{sqdim}_\mu(\mathscr{C})$, is the largest d for which there are d functions $f_1, \ldots, f_d \in \mathscr{C}$ with

$$\left| \mathop{\mathbf{E}}_{x\sim\mu} [f_i(x) \cdot f_j(x)] \right| \leqslant \frac{1}{d}$$

for all $i \neq j$. We denote

$$\mathrm{sqdim}(\mathscr{C}) = \max_\mu \{\mathrm{sqdim}_\mu(\mathscr{C})\}.$$

The SQ dimension is a tight measure [13] of the learning complexity of a given concept class $\mathscr{C}$ in the SQ model. In addition, the SQ dimension is strongly related to complexity theory [32].

3 Approximation Rank: Definition and Properties

For a real matrix A, its ε-*approximation rank* is defined as

$$\mathrm{rank}_\varepsilon(A) = \min_B \{\mathrm{rank}(B) : B \text{ real}, \|A - B\|_\infty \leqslant \varepsilon\}.$$

This notion is a natural one and has been studied before. In particular, Buhrman and de Wolf [4] show that the approximation rank of a matrix implies lower bounds on

its quantum communication complexity (in the bounded-error model without entanglement). In Section 6, we survey two other related concepts: matrix rigidity and dimension complexity.

We define the ε-*approximation rank* of a concept class $\mathscr{C}$ as

$$\mathrm{rank}_\varepsilon(\mathscr{C}) = \mathrm{rank}_\varepsilon(M),$$

where M is the characteristic matrix of $\mathscr{C}$. For example, $\mathrm{rank}_0(\mathscr{C}) = \mathrm{rank}(M)$ and $\mathrm{rank}_1(\mathscr{C}) = 0$. It is thus the behavior of $\mathrm{rank}_\varepsilon(\mathscr{C})$ for intermediate values of ε that is of primary interest. The following proposition follows trivially from our definitions.

Proposition 2 (Approximation rank reinterpreted). *Let $\mathscr{C}$ be a concept class. Then $\mathrm{rank}_\varepsilon(\mathscr{C})$ is the smallest integer r such that there exist real functions $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ with the property that each $f \in \mathscr{C}$ has $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant \varepsilon$ for some reals $\alpha_1, \ldots, \alpha_r$.*

3.1 Improving the Quality of the Approximation

We now take a closer look at $\mathrm{rank}_\varepsilon(M)$ as a function of ε. Suppose we have an estimate of $\mathrm{rank}_E(M)$ for some $0 < E < 1$. Can we use this information to obtain a nontrivial upper bound on $\mathrm{rank}_\varepsilon(M)$, where $0 < \varepsilon < E$? It turns out that we can. We first recall that the sign function can be approximated well by a real polynomial:

Fact 1. *Let $0 < E < 1$ be given. Then for each integer $d \geqslant 1$, there exists a degree-d real univariate polynomial $p(t)$ such that*

$$|p(t) - \mathrm{sign}(t)| \leqslant 8\sqrt{d}\left(1 - \frac{(1-E)^2}{16}\right)^d \qquad (1-E \leqslant |t| \leqslant 1+E).$$

Fact 1 can be extracted with little effort from Rudin's proof [31, Thm. 7.26] of the Weierstrass approximation theorem. Subtler, improved versions of Fact 1 can be readily found in the approximation literature.

Theorem 5. *Let M be a sign matrix, and let $0 < \varepsilon < E < 1$. Then*

$$\mathrm{rank}_\varepsilon(M) \leqslant \mathrm{rank}_E(M)^d,$$

where d is any positive integer with $8\sqrt{d}(1-(1-E)^2/16)^d \leqslant \varepsilon$.

Proof. Let d be as stated. By Fact 1, there is a degree-d polynomial $p(t)$ with

$$|p(t) - \mathrm{sign}(t)| \leqslant \varepsilon \qquad (1-E \leqslant |t| \leqslant 1+E).$$

Let A be a real matrix with $\|A - M\|_\infty \leqslant E$ and $\mathrm{rank}(A) = \mathrm{rank}_E(M)$. Then the matrix $B = [p(A_{ij})]_{i,j}$ approximates M to the desired accuracy: $\|B - M\|_\infty \leqslant \varepsilon$. Since p is a polynomial of degree d, elementary linear algebra shows that $\mathrm{rank}(B) \leqslant \mathrm{rank}(A)^d$. □

Note. The key idea in the proof of Theorem 5 is to improve the quality of the approximating matrix by applying a suitable polynomial to its entries. This idea is not new. For example, Alon [1] uses the same method in the simpler setting of *one-sided* errors.

We will mainly need the following immediate consequences of Theorem 5.

Corollary 1. *Let M be a sign matrix. Let ε, E be constants with $0 < \varepsilon < E < 1$. Then $\mathrm{rank}_\varepsilon(M) \leqslant \mathrm{rank}_E(M)^c$, where $c = c(\varepsilon, E)$ is a constant.*

Corollary 2. *Let M be a sign matrix. Let ε be a constant with $0 < \varepsilon < 1$. Then $\mathrm{rank}_{1/n^c}(M) \leqslant \mathrm{rank}_\varepsilon(M)^{O(\log n)}$ for every constant $c > 0$.*

By Corollary 1, the choice of the constant ε affects $\mathrm{rank}_\varepsilon(M)$ by at most a polynomial factor. When such factors are unimportant, we will adopt $\varepsilon = 1/3$ as a canonical setting.

3.2 Estimating the Approximation Rank

We will use two methods to estimate the approximation rank. The first uses the ε-approximate trace norm of the same matrix, and the second uses its discrepancy.

Lemma 2 (Lower bound via approximate trace norm). *Let $M \in \{-1,1\}^{N \times N}$. Then*

$$\mathrm{rank}_\varepsilon(M) \geqslant \left(\frac{\|M\|_\Sigma^\varepsilon}{(1+\varepsilon)N} \right)^2.$$

Proof. Let A be an arbitrary matrix with $\|M - A\|_\infty \leqslant \varepsilon$. We have:

$$\begin{aligned}(\|M\|_\Sigma^\varepsilon)^2 \leqslant (\|A\|_\Sigma)^2 = \left(\sum_{i=1}^{\mathrm{rank}(A)} \sigma_i(A) \right)^2 &\leqslant \left(\sum_{i=1}^{\mathrm{rank}(A)} \sigma_i(A)^2 \right) \mathrm{rank}(A) \\ = (\|A\|_F)^2 \,\mathrm{rank}(A) &\leqslant (1+\varepsilon)^2 N^2 \,\mathrm{rank}(A).\end{aligned}$$

□

Our second method is as follows.

Lemma 3 (Lower bound via discrepancy). *Let M be a sign matrix and $0 \leqslant \varepsilon < 1$. Then*

$$\mathrm{rank}_\varepsilon(M) \geqslant \frac{1-\varepsilon}{1+\varepsilon} \cdot \frac{1}{64\,\mathrm{disc}(M)^2}.$$

The proof of Lemma 3 requires several definitions and facts that we do not use elsewhere in this paper. For this reason, we defer it to Appendix A.

4 Approximation Rank of Specific Concept Classes

We proceed to prove our main results (Theorems 1–3), restated here as Theorems 7, 9, and 10.

4.1 Disjunctions

We recall a breakthrough result of Razborov [30] on the quantum communication complexity of disjointness. The crux of that work is the following theorem.

Theorem 6 (Razborov [30, Sec. 5.3]). *Let M be the $\binom{n}{n/4} \times \binom{n}{n/4}$ matrix whose rows and columns are indexed by sets in $\binom{[n]}{n/4}$ and entries given by*

$$M_{S,T} = \begin{cases} 1 & \text{if } S \cap T = \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

Then $\|M\|_\Sigma^{1/4} = 2^{\Omega(\sqrt{n})}\binom{n}{n/4}$.

We can now prove an exponential lower bound on the approximation rank of disjunctions, a particularly simple concept class.

Theorem 7 (Approximation rank of disjunctions). *Let $\mathscr{C} = \{\bigvee_{i \in S} x_i : S \subseteq [n]\}$ be the concept class of disjunctions. Then $\operatorname{rank}_{1/3}(\mathscr{C}) = 2^{\Omega(\sqrt{n})}$.*

Proof. One easily verifies that the characteristic matrix of $\mathscr{C}$ is $M_{\mathscr{C}} = [\bigvee_{i=1}^n (x_i \wedge y_i)]_{x,y}$. We can equivalently view $M_{\mathscr{C}}$ as the $2^n \times 2^n$ sign matrix whose rows and columns indexed by sets in $[n]$ and entries given by:

$$M_{\mathscr{C}}(S,T) = \begin{cases} 1 & \text{if } S \cap T = \emptyset, \\ -1 & \text{otherwise.} \end{cases}$$

Now let A be a real matrix with $\|M_{\mathscr{C}} - A\|_\infty \leqslant 1/3$. Let $Z_{\mathscr{C}} = \frac{1}{2}(M_{\mathscr{C}} + J)$, where J is the all-ones matrix. We immediately have $\|Z_{\mathscr{C}} - \frac{1}{2}(A+J)\|_\infty \leqslant 1/6$, and thus

$$\operatorname{rank}_{1/6}(Z_{\mathscr{C}}) \leqslant \operatorname{rank}\left(\tfrac{1}{2}(A+J)\right) \leqslant \operatorname{rank}(A) + 1. \tag{2}$$

However, $Z_{\mathscr{C}}$ contains as a submatrix the matrix M from Theorem 6. Therefore,

$$\operatorname{rank}_{1/6}(Z_{\mathscr{C}}) \;\geqslant \operatorname{rank}_{1/6}(M) \overset{\text{Lem. 2}}{\geqslant} \left(\frac{\|M\|_\Sigma^{1/4}}{(1+1/4)\binom{n}{n/4}}\right)^2 \overset{\text{Thm. 6}}{\geqslant} 2^{\Omega(\sqrt{n})}. \tag{3}$$

The theorem follows immediately from (2) and (3). □

4.2 DNF Formulas

The centerpiece of our proof is the following recent result of Buhrman et al. [5].

Theorem 8 (Buhrman, Vereshchagin, and de Wolf [5, Sec. 3]). *There is a function $f : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ in $\mathsf{AC}^{0,3}$ such that $\operatorname{disc}(f) = 2^{-\Omega(n^{1/3})}$. Moreover, for each fixed y, the function $f_y(x) = f(x,y)$ is a DNF formula of linear size.*

We can now analyze the approximation rank of linear-size DNF formulas.

Theorem 9 (Approximation rank of DNF). *Let $\mathscr{C}$ denote the concept class of functions $f : \{-1,1\}^n \to \{-1,1\}$ computable by DNF formulas of linear size. Then* $\mathrm{rank}_\varepsilon(\mathscr{C}) = 2^{\Omega(n^{1/3})}$ *for* $0 \leqslant \varepsilon \leqslant 1 - 2^{-cn^{1/3}}$, *where $c > 0$ is a sufficiently small absolute constant.*

Proof. Let M be the characteristic matrix of $\mathscr{C}$, and let $f(x,y)$ be the function from Theorem 8. Since $[f(x,y)]_{y,x}$ is a submatrix of M, we have $\mathrm{rank}_\varepsilon(M) \geqslant \mathrm{rank}_\varepsilon([f(x,y)]_{y,x})$. The claim is now immediate from Lemma 3. □

Comparing the results of Theorems 7 and 9 for small constant ε, we see that Theorem 7 is stronger in that it gives a better lower bound against a simpler concept class. On the other hand, Theorem 9 is stronger in that it remains valid for the broad range $0 \leqslant \varepsilon \leqslant 1 - 2^{-\Theta(n^{1/3})}$, whereas the ε-approximation rank in Theorem 7 is easily seen to be at most n for all $\varepsilon \geqslant 1 - \frac{1}{2n}$.

4.3 Majority Functions

As a final application, we consider the concept class $\mathscr{C}$ of majority functions. Here we prove a lower bound of $\Omega(2^n/n)$ on the approximation rank, which is the best of our three constructions.

Theorem 10 (Approximation rank of majority functions). *Let $\mathscr{C}$ denote the concept class of majority functions, $\mathscr{C} = \{\mathrm{MAJ}(\pm x_1, \ldots, \pm x_n)\}$. Then* $\mathrm{rank}_{c/\sqrt{n}}(\mathscr{C}) \geqslant \Omega(2^n/n)$ *for a sufficiently small absolute constant $c > 0$. Also,* $\mathrm{rank}_{1/3}(\mathscr{C}) = 2^{\Omega(n/\log n)}$.

Proof. The characteristic matrix of $\mathscr{C}$ is $M = [\mathrm{MAJ}(x \oplus y)]_{x,y}$. The Fourier spectrum of the majority function has been extensively studied by various authors. In particular, it is well known that

$$\|\widehat{\mathrm{MAJ}}\|_1 = \Omega\left(\frac{2^{n/2}}{\sqrt{n}}\right). \tag{4}$$

(See, e.g., [22, Sec. 7] for a self-contained calculation.) Taking $\varepsilon = c/\sqrt{n}$ for a suitably small constant $c > 0$, we obtain:

$$\mathrm{rank}_{c/\sqrt{n}}(M) \overset{\text{Lem. 2}}{\geqslant} \left(\frac{\|M\|_\Sigma^{c/\sqrt{n}}}{(1+c/\sqrt{n})2^n}\right)^2 \overset{\text{Lem. 1}}{\geqslant} \frac{1}{4}\left(\|\widehat{\mathrm{MAJ}}\|_1 - \frac{c2^{n/2}}{\sqrt{n}}\right)^2 \overset{(4)}{\geqslant} \Omega\left(\frac{2^n}{n}\right).$$

Finally, $\mathrm{rank}_{1/3}(\mathscr{C}) \geqslant [\mathrm{rank}_{c/\sqrt{n}}(\mathscr{C})]^{1/O(\log n)} \geqslant 2^{\Omega(n/\log n)}$ by Corollary 2. □

5 Approximation Rank vs. SQ Dimension

This section relates the approximation rank of a concept class $\mathscr{C}$ to its SQ dimension, a fundamental quantity in learning theory. In short, we prove that (1) the SQ dimension is a lower bound on the approximation rank, and that (2) the gap between the two quantities can be exponential. A starting point in our analysis is the relationship between the SQ dimension of $\mathscr{C}$ and ℓ_2-norm approximation of $\mathscr{C}$, which is also of some independent interest.

Theorem 11 (SQ dimension and ℓ_2 approximation). *Let $\mathscr{C}$ be a concept class, and let μ be a distribution over $\{-1,1\}^n$. Suppose there exist functions $\phi_1,\ldots,\phi_r : \{-1,1\}^n \to \mathbb{R}$ such that each $f \in \mathscr{C}$ has $\mathbf{E}_{x\sim\mu}\left[\left(f(x)-\sum_{i=1}^r \alpha_i\phi_i(x)\right)^2\right] \leqslant \varepsilon$ for some reals $\alpha_1,\ldots,\alpha_r$. Then*

$$r \geqslant (1-\varepsilon)d - \sqrt{d},$$

where $d = \mathrm{sqdim}_\mu(\mathscr{C})$.

Proof. By the definition of the SQ dimension, there exist functions $f_1,\ldots,f_d \in \mathscr{C}$ with $|\mathbf{E}_\mu[f_i \cdot f_j]| \leqslant 1/d$ for all $i \neq j$. For simplicity, assume that μ is a distribution with rational weights (extension to the general case is straightforward). Then there is an integer $k \geqslant 1$ such that each $\mu(x)$ is an integral multiple of $1/k$. Construct the $d \times k$ sign matrix

$$M = [f_i(x)]_{i,x},$$

whose rows are indexed by the functions $f_1,\ldots,f_d$ and whose columns are indexed by inputs $x \in \{-1,1\}^n$ (a given input x indexes exactly $k\mu(x)$ columns). It is easy to verify that $MM^\mathsf{T} = [k\mathbf{E}_\mu[f_i \cdot f_j]]_{i,j}$, and thus

$$\|MM^\mathsf{T} - k\cdot I\|_F < k. \tag{5}$$

The existence of $\phi_1,\ldots,\phi_r$ implies the existence of a rank-r real matrix A with $\|M-A\|_F^2 \leqslant \varepsilon kd$. On the other hand, the Hoffman-Wielandt inequality (Theorem 4) guarantees that $\|M-A\|_F^2 \geqslant \sum_{i=r+1}^d \sigma_i(M)^2$. Combining these two inequalities yields:

$$\begin{aligned}
\varepsilon kd &\geqslant \sum_{i=r+1}^d \sigma_i(M)^2 = \sum_{i=r+1}^d \sigma_i(MM^\mathsf{T}) \\
&\geqslant k(d-r) - \sum_{i=r+1}^d |\sigma_i(MM^\mathsf{T}) - k| \\
&\geqslant k(d-r) - \sqrt{\sum_{i=r+1}^d (\sigma_i(MM^\mathsf{T})-k)^2}\ \sqrt{d-r} && \text{by Cauchy-Swartz} \\
&\geqslant k(d-r) - \|MM^\mathsf{T} - k\cdot I\|_F\ \sqrt{d-r} && \text{by Hoffman-Wielandt} \\
&\geqslant k(d-r) - k\sqrt{d} && \text{by (5).}
\end{aligned}$$

We have shown that $\varepsilon d \geqslant (d-r) - \sqrt{d}$, which is precisely what the theorem claims. To extend the proof to irrational distributions μ, one considers a rational distribution $\tilde{\mu}$ suitably close to μ and repeats the above analysis. We omit these simple details. □

We are now in a position to relate the SQ dimension to the approximation rank.

Theorem 12 (SQ dimension vs. approximation rank). *Let $\mathscr{C}$ be a concept class. Then for $0 \leqslant \varepsilon < 1$,*

$$\mathrm{rank}_\varepsilon(\mathscr{C}) \geqslant (1-\varepsilon^2)\,\mathrm{sqdim}(\mathscr{C}) - \sqrt{\mathrm{sqdim}(\mathscr{C})}. \tag{6}$$

Moreover, there exists a concept class $\mathscr{A}$ with

$$\mathrm{sqdim}(\mathscr{A}) \leqslant O(n^2) \qquad \textit{and} \qquad \mathrm{rank}_{1/3}(\mathscr{A}) \geqslant 2^{\Omega(n/\log n)}.$$

Proof. Let $r = \mathrm{rank}_\varepsilon(\mathscr{C})$. Then there are functions $\phi_1, \ldots, \phi_r$ such that each $f \in \mathscr{C}$ has $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant \varepsilon$ for some reals $\alpha_1, \ldots, \alpha_r$. As a result,

$$\mathbf{E}_\mu \left[(f - \textstyle\sum_{i=1}^r \alpha_i \phi_i)^2 \right] \leqslant \varepsilon^2$$

for every distribution μ. By Theorem 11, $r \geqslant (1-\varepsilon^2)\,\mathrm{sqdim}_\mu(\mathscr{C}) - \sqrt{\mathrm{sqdim}_\mu(\mathscr{C})}$. Maximizing this over μ establishes (6).

To prove the second part, let $\mathscr{A} = \{\mathrm{MAJ}(\pm x_1, \ldots, \pm x_n)\}$. Theorem 10 shows that $\mathscr{A}$ has the stated approximation rank. To bound its SQ dimension, note that each function in $\mathscr{A}$ can be pointwise approximated within error $1 - 1/n$ by a linear combination of the functions $x_1, \ldots, x_n$. Therefore, (6) implies that $\mathrm{sqdim}(\mathscr{A}) \leqslant O(n^2)$. □

6 Related Work

Approximation rank and dimension complexity. Dimension complexity is a fundamental and well-studied notion [7, 8, 22]. It is defined for a sign matrix M as

$$\mathrm{dc}(M) = \min_A \{\mathrm{rank}(A) : A \text{ real}, A_{ij} M_{ij} > 0 \text{ for all } i, j\}.$$

In words, the dimension complexity of M is the smallest rank of a real matrix A that has the same sign pattern as M. Thus, $\mathrm{rank}_\varepsilon(M) \geqslant \mathrm{dc}(M)$ for each sign matrix M and $0 \leqslant \varepsilon < 1$.

Ben-David et al. [2] showed that almost all concept classes with constant VC dimension have dimension complexity $2^{\Omega(n)}$; recall that $\mathrm{dc}(\mathscr{C}) \leqslant 2^n$ always. Forster [7] later developed a powerful tool for lower-bounding the dimension complexity of explicit concept classes. His method has since seen several refinements.

However, this rich body of work is not readily applicable to our problem. Two of the three matrices we study have trivial dimension complexity, and we derive lower bounds on the approximation rank that are exponentially larger. Furthermore, in Theorem 3 we are able to exhibit an explicit concept class with approximation rank $\Omega(2^n/n)$, whereas the highest dimension complexity proved for any explicit concept class is Forster's lower bound of $2^{n/2}$. The key to our results is to bring out, through a variety of techniques, the additional structure in approximation that is not present in sign-representation.

Approximation rank and rigidity. Approximation rank is also closely related to ε-*rigidity*, a variant of matrix rigidity introduced by Lokam [25]. For a fixed real matrix A, its ε-rigidity function is defined as

$$R_A(r, \varepsilon) = \min_B \{\mathrm{weight}(A - B) : \mathrm{rank}(B) \leqslant r,\ \|A - B\|_\infty \leqslant \varepsilon\},$$

where weight$(A-B)$ stands for the number of nonzero entries in $A-B$. In words, $R_A(r,\varepsilon)$ is the minimum number of entries of A that must be perturbed to reduce its rank to r, provided that the perturbation to any single entry is at most ε. We immediately have:

$$\operatorname{rank}_\varepsilon(A) = \min\{r : R_A(r,\varepsilon) \leqslant mn\} \qquad (A \in \mathbb{R}^{m\times n}).$$

As a result, lower bounds on ε-rigidity translate into lower bounds on approximation rank. In particular, ε-rigidity is a more complicated and nuanced quantity. Nontrivial lower bounds on ε-rigidity are known for some special matrix families, most notably the Hadamard matrices [12, 25]. Unfortunately, these results are not applicable to the matrices in our work (see Section 4). To obtain near-optimal lower bounds on approximation rank, we use specialized techniques that target approximation rank without attacking the harder problem of ε-rigidity.

7 Conclusions and Open Problems

This paper studies the ε-approximation rank of a concept class $\mathscr{C}$, defined as the minimum size of a set of features whose linear combinations can pointwise approximate each $f \in \mathscr{C}$ within ε. Our main results give exponential lower bounds on $\operatorname{rank}_\varepsilon(\mathscr{C})$ even for the simplest concept classes. These in turn establish exponential lower bounds on the running time of the known algorithms for distribution-free agnostic learning. An obvious open problem is to develop an approach to agnostic learning that does not rely on pointwise approximation by a small set of features.

Another major open problem is to prove strong lower bounds on the dimension complexity and SQ dimension of natural concept classes. We have shown that

$$\operatorname{rank}_{1/3}(\mathscr{C}) \geqslant \frac{1}{2}\operatorname{sqdim}(\mathscr{C}) - O(1) \qquad \text{and} \qquad \operatorname{rank}_\varepsilon(\mathscr{C}) \geqslant \operatorname{dc}(\mathscr{C}),$$

for each concept class $\mathscr{C}$. In this sense, lower bounds on approximation rank are prerequisites for lower bounds on dimension complexity and the SQ dimension. Of particular interest in this respect are polynomial-size DNF formulas and, more broadly, AC^0 circuits. While this paper obtains strong lower bounds on their approximation rank, it remains a hard open problem to prove an exponential lower bound on their dimension complexity and SQ dimension.

References

1. N. Alon. Problems and results in extremal combinatorics, Part I. *Discrete Mathematics*, 273(1-3):31–53, 2003.
2. S. Ben-David, N. Eiron, and H. U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *J. Mach. Learn. Res.*, 3:441–461, 2003.
3. N. H. Bshouty and C. Tamon. On the Fourier spectrum of monotone functions. *J. ACM*, 43(4):747–770, 1996.
4. H. Buhrman and R. de Wolf. Communication complexity lower bounds by polynomials. In *Conference on Computational Complexity (CCC)*, pages 120–130, 2001.

5. H. Buhrman, N. K. Vereshchagin, and R. de Wolf. On computation and communication with small bias. In *22nd IEEE Conference on Computational Complexity*, 2007.
6. S. E. Decatur. Statistical queries and faulty PAC oracles. In *COLT*, pages 262–268, 1993.
7. J. Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.*, 65(4):612–625, 2002.
8. J. Forster and H. U. Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.*, 350(1):40–48, 2006.
9. G. H. Golub and C. F. V. Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, MD, USA, 3rd edition, 1996.
10. J. C. Jackson. *The harmonic sieve: A novel application of Fourier analysis to machine learning theory and practice*. PhD thesis, Carnegie Mellon University, 1995.
11. A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. In *FOCS: IEEE Symposium on Foundations of Computer Science (FOCS)*, 2005.
12. B. Kashin and A. A. Razborov. Improved lower bounds on the rigidity of Hadamard matrices. *Matematicheskie zametki*, 63(4):535–540, 1998. In Russian.
13. M. Kearns. Efficient noise-tolerant learning from statistical queries. In *STOC '93: Proceedings of the twenty-fifth annual ACM symposium on theory of computing*, pages 392–401, New York, NY, USA, 1993. ACM Press.
14. M. Kearns and M. Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, Aug. 1993.
15. M. J. Kearns, R. E. Shapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.
16. M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, USA, 1994.
17. A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.
18. A. R. Klivans and R. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. In *STOC '01: Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 258–265, New York, NY, USA, 2001. ACM Press.
19. E. Kushilevitz and Y. Mansour. Learning decision trees using the Fourier spectrum. *SIAM J. Comput.*, 22(6):1331–1348, 1993.
20. E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge University Press, New York, NY, USA, 1997.
21. N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform, and learnability. *J. ACM*, 40(3):607–620, 1993.
22. N. Linial, S. Mendelson, G. Schechtman, and A. Shraibman. Complexity measures of sign matrices. *Combinatorica*, 2006. To appear. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/complexity_matrices.ps.gz`.
23. N. Linial and A. Shraibman. Learning complexity vs. communication complexity. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/lcc.pdf`, December 2006.
24. N. Linial and A. Shraibman. Lower bounds in communication complexity based on factorization norms. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/ccfn.pdf`, December 2006.
25. S. V. Lokam. Spectral methods for matrix rigidity with applications to size-depth trade-offs and communication complexity. *J. Comput. Syst. Sci.*, 63(3):449–473, 2001.
26. Y. Mansour. An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. In *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 53–61, New York, NY, USA, 1992. ACM Press.

27. Y. Mansour and M. Parnas. On learning conjunctions with malicious noise. In *ISTCS*, pages 170–175, 1996.
28. R. O'Donnell and R. A. Servedio. Extremal properties of polynomial threshold functions. In *IEEE Conference on Computational Complexity*, pages 3–12, 2003.
29. R. Paturi. On the degree of polynomials that approximate symmetric Boolean functions. In *STOC: ACM Symposium on Theory of Computing (STOC)*, 1992.
30. A. A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya of the Russian Academy of Science, Mathematics*, 67:145–159, 2002.
31. W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 3rd edition, 1976.
32. A. A. Sherstov. Halfspace matrices. In *Proc. of the 22nd Conference on Computational Complexity (CCC)*, 2007.
33. A. A. Sherstov. Separating AC^0 from depth-2 majority circuits. In *Proc. of the 39th Symposium on Theory of Computing (STOC)*, 2007.
34. L. G. Valiant. Learning disjunctions of conjunctions. In *Proc. of the 9th International Joint Conference on Artificial Intelligence, vol. 1*, pages 560–566, Los Angeles, California, 1985.

A Discrepancy and Approximation Rank

The purpose of this section is to prove the relationship between discrepancy and approximation rank needed in Section 4. We start with several definitions and auxiliary results due to Linial et al. [22–24].

For a real matrix A, let $\|A\|_{1\to 2}$ denote the largest Euclidean norm of a column of A, and let $\|A\|_{2\to\infty}$ denote the largest Euclidean norm of a row of A. Define

$$\gamma_2(A) = \min_{XY=A} \|X\|_{2\to\infty}\|Y\|_{1\to 2}.$$

For a sign matrix M, its *margin complexity* is defined as

$$\mathrm{mc}(M) = \min\{\gamma_2(A) : A \text{ real},\ A_{ij}M_{ij} \geqslant 1 \text{ for all } i,j\}.$$

Lemma 4 (Linial et al. [22, Lem. 9]). *Let A be a real matrix. Then $\gamma_2(A) \leqslant \sqrt{\mathrm{rank}(A)\cdot\|A\|_\infty}$.*

Theorem 13 (Linial and Shraibman [23]). *Let M be a sign matrix. Then $\mathrm{mc}(M) \geqslant 1/(8\,\mathrm{disc}(M))$.*

Putting these pieces together yields our desired result:

Lemma 3 (Restated from Sec. 3.2). *Let M be a sign matrix and $0 \leqslant \varepsilon < 1$. Then*

$$\mathrm{rank}_\varepsilon(M) \geqslant \frac{1-\varepsilon}{1+\varepsilon}\cdot\frac{1}{64\,\mathrm{disc}(M)^2}.$$

Proof. Let A be any real matrix with $\|A-M\|_\infty \leqslant \varepsilon$. Put $B = \frac{1}{1-\varepsilon}A$. We have:

$$\begin{aligned}\mathrm{rank}(A) = \mathrm{rank}(B) &\overset{\text{Lem. 4}}{\geqslant} \frac{\gamma_2(B)^2}{\|B\|_\infty} \geqslant \frac{\mathrm{mc}(M)^2}{\|B\|_\infty} \overset{\text{Thm. 13}}{\geqslant} \frac{1}{\|B\|_\infty}\cdot\frac{1}{64\,\mathrm{disc}(M)^2} \\ &\geqslant \frac{1-\varepsilon}{1+\varepsilon}\cdot\frac{1}{64\,\mathrm{disc}(M)^2}.\end{aligned}$$

□