2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)

A New Minimax Theorem for Randomized Algorithms
(Extended Abstract[†])

Shalev Ben-David
Eric Blais
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Canada
(shalev.b | eric.blais)@uwaterloo.ca

***Abstract*—The celebrated minimax principle of Yao (1977) says that for any Boolean-valued function f with finite domain, there is a distribution μ over the domain of f such that computing f to error ϵ against inputs from μ is just as hard as computing f to error ϵ on worst-case inputs. Notably, however, the distribution μ depends on the target error level ϵ: the hard distribution which is tight for bounded error might be trivial to solve to small bias, and the hard distribution which is tight for a small bias level might be far from tight for bounded error levels.**

In this work, we introduce a new type of minimax theorem which can provide a hard distribution μ that works for all bias levels at once. We show that this works for randomized query complexity, randomized communication complexity, some randomized circuit models, quantum query and communication complexities, approximate polynomial degree, and approximate logrank. We also prove an improved version of Impagliazzo's hardcore lemma.

Our proofs rely on two innovations over the classical approach of using Von Neumann's minimax theorem or linear programming duality. First, we use Sion's minimax theorem to prove a minimax theorem for ratios of bilinear functions representing the cost and score of algorithms.

Second, we introduce a new way to analyze low-bias randomized algorithms by viewing them as "forecasting algorithms" evaluated by a certain proper scoring rule. The expected score of the forecasting version of a randomized algorithm appears to be a more fine-grained way of analyzing the bias of the algorithm. We show that such expected scores have many elegant mathematical properties: for example, they can be amplified linearly instead of quadratically. We anticipate forecasting algorithms will find use in future work in which a fine-grained analysis of small-bias algorithms is required.

***Keywords*-Minimax; Randomized computation; Quantum computation; Query complexity; Communication complexity; Polynomial degree complexity; Circuit complexity**

I. Introduction

Yao's minimax principle [Yao77] is a central tool in the analysis of randomized algorithms in many different models of computation. In its most commonly-used form, it states that for every Boolean-valued function f with a finite domain, if $\mathcal{R}^{(c)}$ denotes the set of randomized algorithms with worst-case cost at most c and Δ denotes the set of distributions over the domain of f, then

$$\min_{R\in\mathcal{R}^{(c)}}\max_{\mu\in\Delta}\Pr[R(x)\neq f(x)]=\max_{\mu\in\Delta}\min_{R\in\mathcal{R}^{(c)}}\Pr[R(x)\neq f(x)]$$

with both probabilities being over the choice of x drawn from μ and the internal randomness of R. This identity implies that there exists a distribution μ for which any algorithm that computes f with bounded error over inputs drawn from μ must have cost at least $\mathrm{R}(f)$, the cost of computing f to worst-case bounded error. But it does not say anything else about μ itself. Notably,

I. The minimax principle does not guarantee that the resulting distribution μ must be balanced on the sets $f^{-1}(0)$ and $f^{-1}(1)$.
II. More generally, it does not rule out the possibility that f is very easy to compute by randomized algorithms that are only required to output the correct value with probability at least $\frac{1+\gamma}{2}$ for some small bias measure $\gamma > 0$ over inputs drawn from the distribution μ.

A separate application of the minimax principle can be used to show that there is a distribution μ' for which all randomized algorithms computing f with bias γ over μ' have cost at least $\mathrm{R}_{\frac{1-\gamma}{2}}(f)$ (the cost of computing f to worst-case error $(1-\gamma)/2$), but then there is no guarantee that randomized algorithms with bounded error over μ' must have cost anywhere close to $\mathrm{R}(f)$.

Intuitively, it seems reasonable to expect that for every function f, there is a distribution μ for f that addresses issues I and II: a distribution that is balanced on $f^{-1}(0)$ and $f^{-1}(1)$, and which is at least slightly hard even to solve to a small bias level γ.

Question I.1 (Informal). *Is there a distribution μ which certifies the hardness of f for all bias levels $\gamma > 0$ at the same time?*

More formally, observe that the cost of computing f to worst-case bias γ cannot be smaller than $\gamma^2\,\mathrm{R}(f)$. This is because randomized algorithms can be *amplified*: by repeating an algorithm $O(1/\gamma^2)$ times and outputting the majority vote of the runs, we can increase its bias from γ^2 to $\Omega(1)$. Therefore, a natural refinement of Question I.1 is as follows.

[†] The full version of the article is available on arXiv as report number 2002.10802.

2575-8454/20/$31.00 ©2020 IEEE
DOI 10.1109/FOCS46700.2020.00045

Question I.2 (Refinement of Question I.1)**.** *Is there a distribution μ such that for all bias levels $\gamma > 0$, any algorithm computing f to bias γ against μ must have cost at least $\Omega(\gamma^2 \operatorname{R}(f))$?*

Question I.2 is the primary focus of this work. We answer it affirmatively in a variety of computational models (we can handle most models in which amplification and Yao's minimax principle both apply). We note that the distribution satisfying the conditions of Question I.2 is hard for bounded error in Yao's sense, since each algorithm solving f to bounded error against μ must have cost at least $\Omega(\operatorname{R}(f))$. In addition to this, such μ must also be perfectly balanced between 0- and 1-inputs of f (by considering the limit as $\gamma \to 0$), and must remain somewhat hard to solve even to small bias levels.

The study of Question I.2 has led us to consider randomized forecasting algorithms which output *probabilistic confidence predictions* about the value of $f(x)$, instead of a Boolean guess for $f(x)$. When evaluated using a certain proper scoring rule, the best possible score of a forecasting algorithm is intimately related to the best possible bias of a randomized algorithm; in fact, the score appears to be a more fine-grained way of measuring the bias. Scores of forecasting algorithms appear to be the "right" way of measuring the success of randomized algorithms, as such scores satisfy elegant mathematical properties. The following question, which we answer affirmatively, turns out to be a strengthening of Question I.2.

Question I.3. *Is there a distribution μ such that for all $\eta > 0$, any forecasting algorithm which achieves expected score at least η against μ must have cost at least $\Omega(\eta \operatorname{R}(f))$?*

A. Motivation from joint computation

The answers to Question I.2 and Question I.3 have a direct impact on the study of composition theorems and joint computation problems in randomized computational models: a natural approach for such problems involves first applying a minimax theorem and then establishing the required inequalities in the deterministic distributional setting. However, as observed by Shaltiel [Sha03] this approach runs into trouble if the hard distribution is easy to solve to small bias. Specifically, Shaltiel considered distributions μ which are hard to solve most of the time, but which give a completely trivial input with small probability γ. Then computing n independent copies from μ is a little easier than n times the cost of computing f, because on average, γn of the copies are trivial; the cost of computing n independent inputs from μ is at most $(1-\gamma)n$ times the cost of solving f.

Things get even worse when the inputs have a promised correlation, as can happen when proving composition theorems. For a concrete example, consider the partial function $\textsc{Trivial}_n$, which is defined on domain $\{0^n, 1^n\}$ and maps $0^n \to 0$ and $1^n \to 1$. Suppose we want to prove a composition lower bound with $\textsc{Trivial}_n$ on the outside: that is, we want to show that for every function f, computing $\textsc{Trivial}_n$ composed with n copies of f requires $\Omega(\operatorname{R}(f))$ cost. In other words, we want to lower bound the cost of an algorithm which outputs 0 when given n 0-inputs to f, outputs 1 when given n 1-inputs to f, and outputs arbitrarily when given some other type of input.

Now, if we try to lower bound this using the hard distribution from Yao's minimax principle, then the distribution might give a trivial input with small probability γ, as Shaltiel observed; but then so long as $n = \Omega(1/\gamma)$, one of the inputs to f will be trivial with high probability, and we can solve this "all-0s vs all-1s" problem simply by searching for the trivial copy – potentially much faster than the worst-case cost of computing a single copy of f!

The hard distributions we give in this work solve this issue by being hard for all bias levels. In our companion manuscript [BB20], we use one of the query versions of our minimax theorem to prove a new composition theorem for randomized query complexity.

B. Main tools

Minimax theorem for cost/score ratios: The first main result is a new minimax theorem for the ratio of the cost and score of randomized algorithms. A special case of the theorem with a simple formulation is as follows.

Theorem I.4. *[Special case of the main minimax theorem] Let $\mathcal{R}$ be a set of randomized algorithms that can be expressed as a convex subset of a real topological vector space. Let S be a nonempty finite set, and let Δ be the set of all probability distributions over S, viewed as a subset of $\mathbb{R}^{|S|}$. Let $\operatorname{cost}\colon \mathcal{R} \times \Delta \to (0,\infty)$ and $\operatorname{score}\colon \mathcal{R} \times \Delta \to [-\infty,\infty)$ be continuous bilinear functions. Then using the convention $r/0 = \infty$ for all $r \in (0,\infty)$ and the notation $r^+ = \max\{r, 0\}$ for all $r \in [-\infty,\infty]$, we have*

$$\inf_{R\in\mathcal{R}} \max_{x\in S} \frac{\operatorname{cost}(R,x)}{\operatorname{score}(R,x)^+} = \max_{\mu\in\Delta} \inf_{R\in\mathcal{R}} \frac{\operatorname{cost}(R,\mu)}{\operatorname{score}(R,\mu)^+}.$$

Further, all of the above maximums are attained.

The general version of the minimax theorem in the full version of the paper shows that the same identity holds even when the cost and score functions are semicontinuous and saddle (but not necessarily linear) under some mild additional restrictions. Furthermore, a variant of the theorem also holds when we consider convex and compact subsets of distributions over the finite set S instead of the set Δ of all distributions over that set.

Minimax theorems for ratios of semicontinuous and saddle functions as in the general version of Theorem I.4 do not seem to have appeared in the literature previously in the precise form we need, but as we show in the full version of the paper, they can be obtained by extending Sion's minimax theorem [Sio58] with standard arguments. We believe that the main contribution of the general version of Theorem I.4 is in its interpretation for randomized algorithms. Various extensions and variations of Yao's minimax theorem have been considered in the computer science literature previously [Yao77; Imp95; Ver98; Bra15; BGK+18; BB19], but all of them appear to consider the cost of an algorithm (with the minimax theorem applied to algorithms with a fixed worst-case score), the score of an algorithm (with the cost being fixed), or a linear combination of the two. None of those variants suffice to answer the questions raised at the beginning of the introduction or to establish the results in the following subsections; what was needed in those cases was a minimax theorem for the *ratio* of the cost/score of randomized algorithms, and we suspect that this ratio minimax theorem will find further applications in computer science in the future as well.

Forecasting algorithms and linear amplification: To convert the statements obtained from the general version of Theorem I.4 regarding the cost/score ratios of randomized algorithms under some distribution μ into more familiar lower bounds on the cost of randomized algorithms that achieve some bias on μ, we need a *linear amplification* theorem. Ideally, we would like to argue that if there exists a randomized algorithm R with bias γ on μ, then by combining $O(1/\gamma)$ instances of R we can obtain a randomized algorithm R' with $\mathrm{cost}(R',\mu) = O\left(\frac{1}{\gamma}\cdot\mathrm{cost}(R,\mu)\right) = O\left(\frac{\mathrm{cost}(R,\mu)}{\mathrm{bias}_f(R,\mu)}\right)$ and constant bias. Unfortunately, such a linear amplification property does not hold for most models of randomized algorithms, where amplification from bias γ to bounded error requires combining $O(1/\gamma^2)$ instances of the original algorithm. To obtain a linear amplification result, we must turn our attention away from bias and error and consider other score functions instead.[1]

To describe our score function, we first generalize our computational model from randomized algorithms that output 0 or 1 to *forecasting algorithms*, which are randomized algorithms that output a *confidence value* in $[0,1]$ for the value $f(x)$ of the function f on their given input x. A "low" confidence prediction is a value close to $\frac{1}{2}$ whereas a "high" confidence prediction would be a value very close to 0 or to 1. There are many natural ways to assign a score to a confidence value for $f(x)$. The study of such scoring rules and their properties has a rich history in the statistics and decision theory communities (see for instance [BSS05; GR07] and references therein); we discuss some fundamental scoring rules and give relations between them in the full version of the paper. Of particular importance to our main purpose is the scoring rule $\mathrm{hs}\colon [0,1] \to [-\infty, 1]$ defined by

$$\mathrm{hs}_f(p) = \begin{cases} 1 - \sqrt{\frac{1-p}{p}} & \text{when } f(x) = 1 \\ 1 - \sqrt{\frac{p}{1-p}} & \text{when } f(x) = 0. \end{cases}$$

Define the score of a forecasting algorithm R on an input x in the domain of f to be $\mathrm{score}_{\mathrm{hs},f}(R,x) = \mathbb{E}[\mathrm{hs}_f(R(x))]$, the expectation of the hs score of the output of R over the internal randomness of R. Then linear amplification does hold for this score function.

Lemma I.5. *For any Boolean-valued function f, any forecasting algorithm R, and any $k \geq 1$, there is a forecasting algorithm R' that combines the outputs of k instances of R and satisfies*

$$\mathrm{score}_{\mathrm{hs},f}(R',x) \geq 1 - (1 - \mathrm{score}_{\mathrm{hs},f}(R,x))^k$$

for every x in the domain of f. In particular, when $k = \max_x \frac{2}{\mathrm{score}_{\mathrm{hs},f}(R,x)}$ then for each $x \in \mathrm{Dom}(f)$, $\mathrm{score}_{\mathrm{hs},f}(R',x) \geq 1 - e^{-2} > 0.85$.

To the best of our knowledge, Lemma I.5 has not previously appeared in the literature. This lemma is sensitive to the precise definition of hs_f; other scoring rules do not appear to satisfy this amplification property, which is crucial for the proof of our main results. Additionally, the scoring rule hs_f is special because there is a close connection between hs score of forecasting algorithms and the bias of randomized algorithms.

Lemma I.6. *For any Boolean-valued function f, any distribution μ on $\mathrm{Dom}(f)$, and any parameter $\gamma > 0$,*

- *If there exists a randomized algorithm R with $\mathrm{bias}_f(R,\mu) = 1 - 2\Pr[R(x) \neq f(x)] \geq \gamma$, then there is a forecasting algorithm R' with $\mathrm{score}_{\mathrm{hs},f}(R',\mu) \geq 1 - \sqrt{1-\gamma^2} \geq \gamma^2/2$, and*
- *If there exists a forecasting algorithm R with $\mathrm{score}_{\mathrm{hs},f}(R,\mu) \geq \gamma$ then there is a randomized algorithm R' with $\mathrm{bias}_f(R',\mu) \geq \gamma$.*

Moreover, in both cases R' can be explicitly constructed from R by modifying its output.

Lemma I.5 and Lemma I.6 can be used to reprove the fact that $O(1/\gamma^2)$ instances of a bias-γ randomized algorithms can be combined to obtain a bounded-error algorithm; combining those lemmas (or, more precisely, specific instantiations of these lemmas that account for the explicit constructions of the relevant algorithms and their costs) with the minimax theorem also leads to new results as described in the next section.

C. Main results

Hard distributions for bounded error and small bias: The minimax theorem for cost/score ratios and linear amplification of forecasting algorithms can be combined to show

[1] The astute reader may have noticed that we obtain linear amplification if we simply set the score to be the squared bias of the randomized algorithm. That is true, but this approach does not work in conjunction with the ratio minimax theorem since this score function no longer satisfies the appropriate saddle property requirements of that theorem; this is why we instead consider forecasting algorithms as described below.

Table I

SUMMARY OF RESULTS ON HARD DISTRIBUTIONS FOR BOUNDED ERROR AND SMALL BIAS.

Randomized comm. complexity	$\mathrm{RCC}^{\mu}_{\dot\gamma}(f) = \Omega\big(\gamma^2\, \mathrm{RCC}(f)\big)$
Quantum query complexity	$\mathrm{QDT}^{\mu}_{\dot\gamma}(f) = \gamma \cdot \tilde\Omega\big(\, \mathrm{QDT}(f)\big)$
Quantum comm. complexity	$\mathrm{QCC}^{\mu}_{\dot\gamma}(f) = \gamma \cdot \tilde\Omega\big(\, \mathrm{QCC}(f)\big)$
Polynomial degree	$\deg^{\mu}_{\dot\gamma}(f) = \gamma \cdot \tilde\Omega(\mathrm{adeg}(f))$
Log-rank complexity	$\log \mathrm{rank}^{\mu}_{\dot\gamma}(f) = \gamma \cdot \tilde\Omega(\log \mathrm{rank}_{\frac{1}{3}}(f))$
Circuit complexity	$\mathrm{Rcirc}^{\mu}_{\dot\gamma}(f) = \gamma^2 \cdot \tilde\Omega\big(\, \mathrm{Rcirc}(f)\big)$
Log-depth circuit complexity	$\mathrm{RNC1}^{\mu}_{\dot\gamma}(f) = \gamma^2 \cdot \tilde\Omega\big(\, \mathrm{RNC1}(f)\big)$
Threshold circuit complexity	$\mathrm{RTC0}^{\mu}_{\dot\gamma}(f) = \gamma^2 \cdot \tilde\Omega\big(\, \mathrm{RTC0}(f)\big)$

that for many measures of randomized complexity, for every Boolean-valued function f with finite domain there exists a single distribution μ on which it is hard to compute f with bounded error *and* with (any) small bias. For example, letting $\mathrm{RDT}(f)$ denote the minimum (worst-case) query complexity of a randomized algorithm computing f (or equivalently the minimum worst-case depth of a decision tree computing f) with error at most $\frac{1}{3}$ on every input in $\mathrm{Dom}(f)$ and $\mathrm{RDT}^{\mu}_{\dot\gamma}$ denote the minimum query complexity of a randomized algorithm that has error probability at most $\dot\gamma := \frac{1-\gamma}{2}$ when inputs are drawn from μ, we obtain the following result.

Theorem I.7. *For any non-constant partial function $f\colon \{0,1\}^n \to \{0,1\}$, there exists a distribution μ on $\mathrm{Dom}(f)$ such that for every $\gamma \in [0,1]$,*

$$\mathrm{RDT}^{\mu}_{\dot\gamma}(f) = \Omega\big(\gamma^2\, \mathrm{RDT}(f)\big).$$

We establish analogous theorems for multiple other computational models as well; see Table I. (Note that as in Theorem I.7, the novel aspect of all these results is that they guarantee that for each of the stated inequalities, there exists a *single* distribution μ that satisfies the inequality for *every* value of γ simultaneously.)

Hard distributions for forecasting algorithms: The theorems listed above settle Question I.2 in the affirmative for the specified models. For the models with quadratic dependence on γ (i.e. randomized query complexity, randomized communication complexity, and the various randomized circuit models), we also get hard distributions which lower bound the expected score of a forecasting algorithm, settling Question I.3 affirmatively.

Distinguishing power of randomized algorithms and protocols: In the communication complexity setting, we can also analyze how well a randomized communication protocol computes a function $f\colon \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ via its communication transcripts. Let $\mathrm{tran}(R, \mu_0)$ denote the distribution on communication transcripts of the randomized protocol R on inputs drawn from μ. Then one way to measure how well R is able to distinguish 0- and 1-inputs of f is to measure the Hellinger distance between the distributions $\mathrm{tran}(R, \mu_0)$ and $\mathrm{tran}(R, \mu_1)$ of transcripts of R on some distributions μ_0 over $f^{-1}(0)$ and μ_1 over $f^{-1}(1)$. We can use the minimax and linear amplification theorems to give a strong upper bound on this Hellinger distance as a measure of the cost of the protocol.

Theorem I.8. *For any non-constant partial function $f\colon \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ over finite sets $\mathcal{X}$ and $\mathcal{Y}$, there is a pair of distributions μ_0 on $f^{-1}(0)$ and μ_1 on $f^{-1}(1)$ such that for any randomized communication protocol R, the squared Hellinger distance between the distribution of its transcripts on μ_0 and μ_1 is bounded above by*

$$\mathrm{h}^2\left(\mathrm{tran}(R, \mu_0), \mathrm{tran}(R, \mu_1)\right) \\ = O\left(\frac{\min\{\mathrm{cost}(R, \mu_0), \mathrm{cost}(R, \mu_1)\}}{\mathrm{RCC}(f)}\right).$$

Here $\mathrm{cost}(R, \mu)$ denotes the expected amount of communication the protocol R transmits when given inputs from μ.

The full version of the paper includes an analogous result for query complexity. In our companion paper [BB20], that theorem is one of the ingredients that enables us to establish a new composition theory for query complexity.

Hardcore lemma: Impagliazzo's Hardcore Lemma [Imp95] states that for every $\epsilon, \delta > 0$, if every circuit C of size at most s computes f with error at least δ on the uniform distribution, then there is a δ-regular distribution $\mu = \mu(\delta, \epsilon)$ for which every circuit that computes f with bias at least ϵ on the distribution μ must have size $\Omega(\epsilon^2 s)$. Informally, the lemma shows that if a function f is mildly hard on average, it is because it is "very" hard to compute on a fairly large subset of its inputs. But, interestingly, this version of the hardcore lemma leaves open the possibility that the hard core might be different for various levels ϵ of hardness. Using our main theorems, we can show that this is not the case.

Theorem I.9. *There exists a universal constant $c > 0$ such that for any $\delta > 0$ and function $f : \{0,1\}^n \to \{0,1\}$, if every circuit C of size at most s satisfies $\Pr[C(x) = f(x)] \le 1-\delta$ when the probability is taken over the uniform distribution of x in $\{0,1\}^n$, then there is a distribution μ with min-entropy δ such that for every $\epsilon > 0$, any circuit C' of size at most $c \cdot \epsilon^2/\log(1/\delta) \cdot s$ has success probability bounded by*

$$\Pr[C'(x) = f(x)] \le \frac{1+\epsilon}{2}.$$

The proof of Theorem I.9 follows closely the original argument of Nisan in [Imp95] that established the hardcore lemma via a minimax theorem. Since that original work, many extensions and different proofs of the hardcore lemma have been established (e.g., [Imp95; KS03; BHK09; TTV09]), but to the best of our knowledge Theorem I.9 represents the first version of the lemma which gives a single distribution μ which is hard for all values of $\epsilon > 0$ simultaneously.

D. Recent independent work

In independent work concurrent with this one, Bassilakis, Drucker, Göös, Hu, Ma, and Tan [BDG+20] showed the existence of a certain hard distribution for randomized query complexity. They showed every Boolean function f has hard distributions μ_0 and μ_1 (on 0- and 1-inputs respectively) such that given query access to k independent samples from μ_b, it is still necessary to use $\Omega(\mathrm{R}(f))$ queries to the bits of the samples in order to decide the value of $b \in \{0,1\}$ to bounded error.

The guarantee on the hard distribution provided by [BDG+20] is formally stronger than the one we provide in the query complexity analogue of Theorem I.8 in the full version of the paper (though in our companion manuscript [BB20], we prove a new composition theorem for randomized query complexity, and use it to conclude that the guarantee of [BDG+20] turns out to be equivalent to the guarantee of the query complexity analogue of Theorem I.8). The tools used by [BDG+20] are also completely different: they use arguments specific to query complexity that construct the hard distribution more explicitly, but their arguments do not generalize to other models such as communication complexity or circuit complexity.

E. Overview of the remaining sections of the full version

Section II is devoted to proving the main minimax theorem for the cost/score ratio of randomized algorithms. The main result of that section is the general version of Theorem I.4; the rest of the section is devoted to introducing the mathematical notions and preliminaries required to obtain a proof of that theorem from Sion's minimax theorem.

Section III introduces the basic definitions and some basic scoring rules for forecasting algorithms. The section establishes some of the core properties of scoring functions, including notably connections between the best score achievable by forecasting algorithms on distributions over inputs and various distance measures on those distributions. The final portions of this section then establish the main linear amplification theorem in general form in Lemma I.5 and the general form of the conversion between randomized and forecasting algorithms.

Section IV focuses on the query and communication complexity settings. Conversions between randomized and forecasting algorithms in the query complexity setting are straightforward, but there is one significant challenge in applying the linear amplification theorem to obtain the results in Theorem I.7 and its query complexity analogue: the cost and score of a randomized algorithm R on an input x can both depend on x itself. This is a problem because to obtain a constant score (and after the final conversion, a bounded-error randomized algorithm), we want to amplify R with a number k of copies that depends on the score of R on x—but since we don't know x we don't know what $\mathrm{score}(R,x)$ is either. We get around this problem with odometer arguments: by empirically estimating the expected number of queries R makes on x, we can obtain effective bounds on the number k of copies of R that we need to obtain successful amplification.

As we show in the section, the communication complexity results follow immediately from their query complexity analogues.

Section V establishes the results in the quantum query and communication complexity settings. Unlike in the classical setting, amplification that is linear in the bias of an algorithm *does* hold in the quantum query complexity setting. However, the proof of the minimax theorem for quantum query complexity requires that the set of algorithms must be representable as a convex subset of a real topological space, and that the cost of an algorithm is a convex function on this set. It is not immediately clear how quantum query algorithms can satisfy this condition, because in the usual definition, the cost of a mixture of two quantum algorithms would be the *maximum* of the costs of the algorithms rather than the average. To overcome this issue, we instead establish the main theorem via consideration of what we call *probabilistic* quantum algorithms, which correspond to probability distributions over quantum algorithms and do easily satisfy the appropriate convexity requirements. Probabilistic quantum algorithms are harder to amplify than regular quantum algorithms (due to their lack of coherence), but we show that a linear amplification theorem still holds.

Another important difference between the quantum and the classical setting is that the communication complexity result is not implied by the analogous query complexity result. Nonetheless, the same argument used for quantum query algorithms also holds for quantum communication protocols as well. We complete the proof of the minimax theorem for quantum communication complexity by first providing an abstraction of the query complexity argument and then showing how communication protocols satisfy the conditions of this abstract theorem.

Section VI considers the approximate polynomial degree and the logrank complexity of functions. As with quantum query complexity, approximate polynomial degree satisfies an amplification theorem that is linear in the bias, meaning that we do not need to use forecasting algorithms or scoring rules. However, also as with quantum query complexity, polynomials and their cost do not satisfy the right convexity requirements, as the degree of a mixture of two polynomials is not the average of their degrees. We overcome this by considering probabilistic polynomials. Proving an amplification theorem for probabilistic polynomials turns out to be somewhat tricky, and requires tools from approximation theory such as Jackson's theorem.

Approximate logrank inherits all of the problems of approximate polynomial degree, and adds a few more. To handle approximate logrank, we switch over to the nearly-equivalent model of the logarithm of the approximate gamma

2 norm, and then use the previous trick of considering the *probabilistic* approximate gamma 2 norm. To prove an amplification theorem for probabilistic gamma 2 norm we apply the same tools as for probabilistic polynomials.

Section VII establishes the circuit complexity results. There are two main hurdles in establishing the minimax theorem for randomized circuit complexity. The first is that the notion of randomized circuits is not as trivially extendable to forecasting circuits as in other computational models. We show that this conversion can be done efficiently when we discretize the set of confidence values that can be returned by forecasting circuits, and that this discretization does not affect the guaranteed relations between score and bias. The second is that the overhead required to combine the output of multiple instances of a randomized circuit during linear amplification is not trivial. This second hurdle can be overcome with the use of efficient circuit constructions for elementary arithmetic operations and the iterated addition problem.

The proof of the universal hardcore lemma is obtained via a slight generalization of the ratio minimax theorem.

F. Further remarks and open problems

We make a few remarks regarding other possible generalizations of Yao's original minimax theorem. First, one may wonder why we provide a hard distribution μ satisfying $\mathrm{R}^{\mu}_{\dot\gamma}(f) = \Omega(\gamma^2 \mathrm{R}(f))$ for all γ, rather than the stronger statement $\mathrm{R}^{\mu}_{\dot\gamma}(f) = \Omega(\mathrm{R}_{\dot\gamma}(f))$ for all γ. In other words, we've stated our lower bounds in terms of the bounded-error randomized cost $\mathrm{R}(f)$, which required amplification; why not directly compare the average-case complexity to bias γ, denoted $\mathrm{R}^{\mu}_{\dot\gamma}(f)$, to the worst-case complexity to bias γ, denoted $\mathrm{R}_{\dot\gamma}(f)$?

The reason is that this stronger version of the minimax is actually false: that is, there need not be a distribution μ for which $\mathrm{R}^{\mu}_{\dot\gamma}(f) = \Omega(\mathrm{R}_{\dot\gamma}(f))$ for all γ (even though for every given γ, such a distribution μ that depends on γ does exist, by Yao's minimax theorem). For a counterexample, consider the query complexity model. Let f be the Boolean function on $n+m+1$ bits, where if the first bit is 0 the function f evaluates to the parity of the next m bits, whereas if the first bit is 1 the function f evaluates to the majority of the last n bits. Say we take $n = m^2$. Then, since parity is hard to compute even to small bias, we have $\mathrm{R}_{\dot\gamma}(f) \geq m$ for all γ. We also have $\mathrm{R}_{1/3}(f) = \Omega(m^2)$, since majority on m^2 bits requires $\Omega(m^2)$ queries. Now, consider any distribution μ over the domain of f. If μ places nonzero probability mass on inputs with first bit 1, then μ can necessarily be solved to some sufficiently small bias using at most 2 queries (one query to the first bit of the input, and one to a random position in the input to majority). In this case, we would have $\mathrm{R}^{\mu}_{\dot\gamma}(f) = O(1)$ and $\mathrm{R}_{\dot\gamma}(f) = \Omega(\sqrt{n})$ for this sufficiently small γ. Alternatively, if μ places zero probability mass on inputs with first bit 1, then solving f against μ is solving parity on $m = O(\sqrt{n})$ bits; hence $\mathrm{R}^{\mu}_{1/3}(f) = O(\sqrt{n})$, even though $\mathrm{R}_{1/3}(f) = \Omega(n)$. Similar counterexamples can be constructed in other computational models.

Another possible generalization of Yao's minimax is to a distribution μ for which $\mathrm{R}^{\mu}(f)$ is large even when the both the error of the algorithm and the expected cost are measured against μ. That is, in a normal application of Yao's minimax, we either consider randomized algorithms which only ever make at most T queries (against any input) and measure their expected error against μ, or else we consider randomized algorithms which only ever make error at most ϵ (against any input) and measure their expected cost against μ. One may wonder if it is possible for one distribution to certify the hardness of f in both ways at once, with both the cost and the error measured in expectation against μ.

The answer turns out to be yes, as first observed by Vereshchagin for query complexity [Ver98]. Vereshchagin stated his theorem for bounded error, but in the case of small bias γ, his techniques appear to give a distribution μ (which depends on γ) such that $\mathrm{R}^{\mu}_{\dot\gamma}(f) = \Omega(\gamma \mathrm{R}_{\dot\gamma}(f))$ even where the left-hand side is defined as the *expected* query complexity against μ to bias at least γ (also against μ). This is in contrast to Yao-style minimax theorems, which are stronger in that they lack the γ factor on the right hand side, but weaker in that the left-hand side has either the cost or the error being worst-case (rather than both being average-case against μ).

Our results in this work are "Vereshchagin-like" in that they hold even when $\mathrm{R}^{\mu}_{\dot\gamma}(f)$ has both the cost and the bias defined in expectation against μ. We prove such results for randomized query complexity and randomized communication complexity, showing a single μ satisfies $\mathrm{R}^{\mu}_{\dot\gamma}(f) = \Omega(\gamma^2 \mathrm{R}(f))$ for all $\gamma > 0$, even when both the error and the cost in the definition of $\mathrm{R}^{\mu}_{\dot\gamma}(f)$ are average-case against μ. (For models such as quantum query complexity or circuit complexity, the expected cost of an algorithm does not have an obvious interpretation, since the algorithms generally have the same cost for all inputs; therefore, for those models we do not give a theorem in which the cost is measured in expectation against μ.)

Note that our minimax theorem is not directly comparable to Vereshagin, because we state our lower bounds in an "amplified" form – that is, the lower bounds are with respect to $\mathrm{R}(f)$ rather than $\mathrm{R}_{\dot\gamma}(f)$. As previously mentioned, this is necessary when proving that a single distribution works for all γ, and our theorems appear to be tight in that setting. Moreover, Vereshchagin's theorem is tight in its setting: the factor of γ is necessary, because average-case query complexity can be smaller than worst-case query complexity (for example, consider the parity function on n bits, which has $\mathrm{R}_{\dot\gamma}(f) = n$ for all γ; if we design a randomized algorithm which queries all the bits with probability γ and queries no bits with probability $1-\gamma$, it will use only γn

expected queries, and it will solve f to bias γ).[2]

A remaining open problem is as follows: can Vereshchagin's theorem be modified to show

$$\mathrm{R}^{\mu}_{\dot{\gamma}}(f) = \Omega(\overline{\mathrm{R}}_{\dot{\gamma}}(f)), \tag{1}$$

where both cost and bias on the left are measured in expectation against μ, and where $\overline{\mathrm{R}}_{\dot{\gamma}}(f)$ denotes the worst-case (over the inputs of f) expected (over the internal randomness of the algorithm) query complexity of f to bias γ? Note that in the bounded-error setting, $\overline{\mathrm{R}}(f) = \Theta(\mathrm{R}(f))$, so for bounded γ this result follows from both Vereshchagin's theorem and from our work here. For small γ, we leave this question as an intriguing open problem.

We also note that we cannot hope that a single distribution μ satisfies (1) for all γ, because one can construct a counterexample via a modification of our earlier function: we let f be defined on $1 + m + n$ bits, where if $x_1 = 0$ the function evaluates to the parity of the next m bits, and if $x_1 = 1$ the function evaluates to the majority of the last n bits, as before; this time we will have $n = m^{4/3}$. We also add a promise: we require that the input always has Hamming weight either at most $n/2 - \sqrt{n}$ or at least $n/2 + \sqrt{n}$ on the last n bits, turning the majority part of the function into a $\sqrt{n}$-gap majority function. Now, to compute f to worst-case bias γ requires at least γm expected queries on inputs x with $x_1 = 0$, and requires at least $\gamma^2 n$ expected queries on inputs with $x_1 = 1$, so at least $\Omega(\max\{\gamma m, \gamma^2 n\})$ expected queries in the worst case. This is $\Omega(n^{1/4})$ when $\gamma = n^{-1/2}$ and $\Omega(n)$ when γ is constant. Now fix a distribution μ, let p be the probability that μ assigns to inputs with $x_1 = 1$. If $p \leq 1/2$, then we can compute f to constant bias simply by querying the first bit, guessing randomly if $x_1 = 1$, and querying m bits to compute f exactly when $x_1 = 0$; this uses $O(n^{3/4})$ queries to achieve constant bias, instead of the $\Omega(n)$ which were required in the worst case. On the other hand, if $p \geq 1/2$, then we can compute f against μ by querying the first bit and nothing else when $x_1 = 0$ (guessing the answer randomly), and otherwise making one additional query to estimate the gap majority function to bias $1/\sqrt{n}$. This uses 2 queries and achieves bias $n^{-1/2}$ against μ, instead of the $\Omega(n^{1/4})$ queries required in the worst case.

Acknowledgements

We thank Justin Thaler for discussions and references related to approximate polynomial degree and its amplification. We also thank Andrew Drucker, Mika Göös, and Li-Yang Tan for correspondence about their ongoing work [BDG+20]. We thank anonymous reviewers for many helpful comments.

References

[Alt88] Helmut Alt. Comparing the combinational complexities of arithmetic functions. *Journal of the ACM* (1988). DOI: 10.1145/42282.214084.

[AR20] Scott Aaronson and Patrick Rall. Quantum Approximate Counting, Simplified. *Proceedings of the 3rd Symposium on Simplicity in Algorithms (SOSA)*. 2020. DOI: 10.1137/1.9781611976014.5. arXiv: 1908.10846.

[BB19] Eric Blais and Joshua Brody. Optimal Separation and Strong Direct Sum for Randomized Query Complexity. *Proceedings of the 34th Conference on Computational Complexity (CCC)*. 2019. DOI: 10.4230/LIPICS.CCC.2019.29. arXiv: 1908.01020.

[BB20] Shalev Ben-David and Eric Blais. A tight composition theorem for the randomized query complexity of partial functions. *Proceedings of the 61st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2020. arXiv: 2002.10809.

[BBGK18] Shalev Ben-David, Adam Bouland, Ankit Garg, and Robin Kothari. Classical Lower Bounds from Quantum Upper Bounds. *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 2018. DOI: 10.1109/focs.2018.00040. arXiv: 1807.06256.

[BCH86] Paul W. Beame, Stephen A. Cook, and H. James Hoover. Log Depth Circuits for Division and Related Problems. *SIAM Journal on Computing* (1986). Previous version in FOCS 1984. DOI: 10.1137/0215070.

[BDG+20] Andrew Bassilakis, Andrew Drucker, Mika Göös, Lunjia Hu, Weiyun Ma, and Li-Yang Tan. The Power of Many Samples in Query Complexity. *Proceedings of the 47th International Colloquium on Automata, Languages, and Programming (ICALP)*. 2020. DOI: 10.4230/LIPIcs.ICALP.2020.9. arXiv: 2002.10654.

[BGK+18] Mark Braverman, Ankit Garg, Young Kun Ko, Jieming Mao, and Dave Touchette. Near-Optimal Bounds on the Bounded-Round Quantum Communication Complexity of Disjointness. *SIAM Journal on Computing* (2018). Previous version in FOCS 2015. DOI: 10.1137/16m1061400. arXiv: 1505.03110.

[BHK09] Boaz Barak, Moritz Hardt, and Satyen Kale. The Uniform Hardcore Lemma via Approximate Bregman Projections. *Proceedings of the 20th Annual ACM-SIAM Symposium on*

[2] We thank an anonymous reviewer for this example.

Discrete Algorithms. 2009. DOI: 10.1137/1.9781611973068.129.

[BHMT02] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. *Proceedings of an AMS Special Session on Quantum Computation and Information (CONM)*. 2002. DOI: 10.1090/conm/305/05215. arXiv: quant-ph/0005055.

[BNRW07] Harry Buhrman, Ilan Newman, Hein Rohrig, and Ronald de Wolf. Robust Polynomials and Quantum Algorithms. *Theory of Computing Systems* (2007). Previous version in STACS 2005. DOI: 10.1007/s00224-006-1313-z. arXiv: quant-ph/0309220.

[Bra15] Mark Braverman. Interactive Information Complexity. *SIAM Journal on Computing* (2015). Previous version in STOC 2012. DOI: 10.1137/130938517.

[BSS05] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Preprint, 2005. URL: pdfs.semanticscholar.org/d670/6b6e626c15680688b0774419662f2341caee.pdf.

[CSV84] Ashok K. Chandra, Larry Stockmeyer, and Uzi Vishkin. Constant Depth Reducibility. *SIAM Journal on Computing* (1984). DOI: 10.1137/0213028.

[GKKT17] Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably Learning the ReLU in Polynomial Time. *Proceedings of the 30th Annual Conference on Learning Theory (COLT)*. 2017. arXiv: 1611.10258.

[GR07] Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* (2007). DOI: 10.1198/016214506000001437.

[Imp95] R. Impagliazzo. Hard-core distributions for somewhat hard problems. *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 1995. DOI: 10.1109/sfcs.1995.492584.

[Jac11] Dunham Jackson. “Über die Genauigkeit der Annäherung stetiger Funktionen durch ganze rationale Funktionen gegebenen Grades und trigonometrische Summen gegebener Ordnung”. PhD thesis. University of Göttingen, 1911. URL: gdz.sub.uni-goettingen.de/id/PPN30230648X.

[KS03] Adam R. Klivans and Rocco A. Servedio. Boosting and Hard-Core Set Construction. *Machine Learning* (2003). Previous version in FOCS 1999. DOI: 10.1023/a:1022949332276.

[LS09] Troy Lee and Adi Shraibman. An Approximation Algorithm for Approximation Rank. *Proceedings of the 24th Conference on Computational Complexity (CCC)*. 2009. DOI: 10.1109/ccc.2009.25. arXiv: 0809.2093.

[LSŠ08] Troy Lee, Adi Shraibman, and Robert Špalek. A Direct Product Theorem for Discrepancy. *Proceedings of the 23rd Conference on Computational Complexity (CCC)*. 2008. DOI: 10.1109/ccc.2008.25.

[MCAL17] Marianthi Markatou, Yang Chen, Georgios Afendras, and Bruce G. Lindsay. Statistical Distances and Their Role in Robustness. *New Advances in Statistics and Data Science* (2017). DOI: 10.1007/978-3-319-69416-0_1. arXiv: 1612.07408.

[MMR94] G. V. Milovanovic, D. S. Mitrinovic, and Th. M. Rassias. *Topics in Polynomials: Extremal Problems, Inequalities, Zeros*. World Scientific, 1994. ISBN: 978-981-02-0499-0. DOI: 10.1142/1284.

[Ofm62] Yuri P. Ofman. On the algorithmic complexity of discrete functions. *Doklady Akademii Nauk* (1962).

[Pip87] Nicholas Pippenger. The complexity of computations by networks. *IBM Journal of Research and Development* (1987). DOI: 10.1147/rd.312.0235.

[RT92] John H. Reif and Stephen R. Tate. On Threshold Circuits and Polynomial Computation. *SIAM Journal on Computing* (1992). DOI: 10.1137/0221053.

[Sha03] Ronen Shaltiel. Towards proving strong direct product theorems. *Computational Complexity* (2003). Previous version in CCC 2001. DOI: 10.1007/s00037-003-0175-x. ECCC: 2001/009.

[She12] Alexander A. Sherstov. Strong Direct Product Theorems for Quantum Communication and Query Complexity. *SIAM Journal on Computing* (2012). Previous version in STOC 2011. DOI: 10.1137/110842661. arXiv: 1011.4935.

[She13] Alexander A. Sherstov. Making Polynomials Robust to Noise. *Theory of Computing* (2013). Previous version in STOC 2012. DOI: 10.4086/toc.2013.v009a018. ECCC: 2012/037.

[Sio58] Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics* (1958). DOI: 10.2140/pjm.1958.8.171.

[Tøp00] Flemming Tøpsoe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory* (2000). DOI: 10.1109/18.850703.

[TTV09] Luca Trevisan, Madhur Tulsiani, and Salil Vadhan. Regularity, Boosting, and Efficiently Simulating Every High-Entropy Distribution. *Proceedings of the 24th Conference on Computational Complexity (CCC)*. 2009. DOI: 10.1109/ccc.2009.41. ECCC: 2008/103.

[Ver98] Nikolai K. Vereshchagin. Randomized Boolean decision trees: Several remarks. *Theoretical Computer Science* (1998). DOI: 10.1016/S0304-3975(98)00071-1.

[Vol99] Heribert Vollmer. *Introduction to Circuit Complexity: A Uniform Approach*. Springer Berlin Heidelberg, 1999. ISBN: 978-3-642-08398-3. DOI: 10.1007/978-3-662-03927-4.

[Weg87] Ingo Wegener. *The Complexity of Boolean Functions*. Wiley, 1987. ISBN: 3-519-02107-2. URL: eccc.weizmann.ac.il/static/books/The_Complexity_of_Boolean_Functions/.

[Yao77] Andrew Yao. Probabilistic computations: toward a unified measure of complexity. *Proceedings of the 18th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (1977). DOI: 10.1109/SFCS.1977.24.