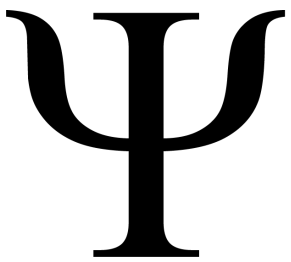


Quantum-Classical Gaps and Quantum Shallow Circuits

Matthew Fox

An essay submitted
for partial fulfillment of
Perimeter Scholars International

June, 2022

Contents

Quantum-Classical Gaps and Quantum Shallow Circuits

Matthew Fox

Supervisor: Robert Spekkens

Quantum mechanics (QM) can be operationally construed as a generalized probability theory based on the theory of Hilbert spaces. In this view, it is natural to wonder if QM is in fact different from classical probability theory, or if it is just a syntactically distinct but nevertheless probabilistically identical formalism.

That QM does indeed transcend classical probability theory follows from Bell's eponymous theorem, which shows that under certain conditional independence assumptions there are strictly more correlations achievable in QM than in any classical theory. This fact can be cast in terms of a quantum-classical gap between the distributions that are compatible with particular generalized Bayesian networks.

In this essay we attempt to understand the advantage of quantum shallow circuits in terms of quantum-classical gaps. While we are unable to prove an advantage in this formalism, we are able to relate generalized Bayesian networks to quantum shallow circuits.

Statement of Original Research

This essay is mostly a literature review, although it does contain original research within Sections 2 and 3 as well as Appendices A and B.

Introduction

> *"We can give up on our rule about what the computer was, we can say: Let the computer itself be built of quantum mechanical elements which obey quantum mechanical laws."*
>
> ∼ Richard Feynman [1]

Computation is a physical process. As such, the laws of physics constrain what is computable. If, for example, the laws of physics constitute a consistent and effectively generated formal system, then any model thereof will contain truths not provable by physical devices [2, 3].

In addition, the laws of physics constrain how efficiently a computation can be done. If, for example, acausal structures like Deutschian closed timelike curves exist [4], then determining if the Riemann Hypothesis has a proof in under N lines is child's play [5].

To date, our most promising view of the fundamental world is afforded by quantum mechanics. Thus, quantum mechanics ought to constrain what and how quickly computational tasks can be done. When it comes to query complexity, quantum computers offer a provable computational advantage: a quantum computer can find a needle among N haystacks in $\Theta(\sqrt{N})$ queries, whereas the best classical computers require $\Theta(N)$ queries. Therefore, quantum computers "square-root" the time it takes to complete any task that is constant-time reducible to an unstructured search [6].

However, this speedup is polynomial in the input size, and is also with respect to the query model of computation. Might quantum computers afford an *exponential* speedup with no black boxes? This, and nearly all complexity theoretic questions of comparable flavor, is not known.

That said, there is reason to believe exponential speedups are possible, at least for certain computational tasks. Shor's factoring algorithm is the paradigmatic case in point, which proves that a quantum computer can find a prime factor of an odd and composite N-bit number in just $O(N^3)$ computational steps [7]. Since it is generally believed (though unproven) that factoring on a classical computer is hard, it is generally believed (though equally unproven) that quantum computers afford an exponential speedup.

The main outstanding question that motivated this essay is the following: if quantum computers afford a computational speedup, then why? Of course, it is plain that the answer has something to do with quantum mechanics, but what? To make any progress on this question it is necessary to

scrutinize what makes a physical theory "quantum" in the first place. What we call the *canonically quantum* phenomena are familiar but misguided examples: superposition, entanglement, teleportation, no-cloning, etc. Indeed, the Gottesman-Knill theorem proves that these alone are insufficient for a quantum advantage [8]. Therefore, an advantage must stem from something *quintessentially quantum*, like contextuality [9] or the violation of a Bell inequality [10].

In this essay, we attempt to understand the proven advantage of quantum shallow circuits using generalized Bayesian network [9, 11, 12]. Such a formalism cuts straight through the metaphysical quandaries that the canonically quantum phenomena present, and instead deals purely with the probabilistic correlations of the theory [13–15]. While we are ultimately unable to prove an advantage using this formalism, we are able to relate generalized Bayesian networks to quantum shallow circuits.

This essay covers several topics from Bayesian networks to complexity theory to one-way quantum computing. The first chapter reviews several elementary notions from graph theory, probability theory, the theory of Bayesian networks, and the theory of semialgebraic statistics. The second chapter explores how the theory of Bayesian networks can be generalized to quantum mechanics. Here we introduce the all-important concept of a "quantum-classical gap". The third and final chapter discusses basic computability and complexity theory, the circuit based models of classical and quantum computation, the one-way quantum computer, and our definitions of particular shallow circuit complexity classes. This chapter culminates with a discussion of the quantum advantage of shallow circuits and its relation to quantum-classical gaps.

All proofs in this essay are either in Appendix B or in the reference(s) preceding the claim.

1 Mathematical Preliminaries

Here we review several basic notions from graph theory, probability theory, the theory of Bayesian networks, and the theory of semialgebraic statistics that are used ubiquitously throughout this essay.

1.1 Graph Theory

Let $G = (V, E)$ be an undirected graph with vertex set V and edge set $E \subseteq V \times V$. The *degree* of a vertex $v \in V$, denoted $\deg(v)$, is the number of edges containing v: $\deg(v) = |\{\{u, v\} \in E \mid u \in V\}|$. It is easy to prove that

$$\sum_{v \in V} \deg(v) = 2|E|. \tag{1}$$

This equation is sometimes called the *handshaking lemma* [16].

An *edge coloring* of an undirected graph $G = (V, E)$ is a map $c : E \to \{1, \ldots, \chi_c\}$ satisfying

$$c(e) = c(e') \iff e \cap e' = \varnothing. \tag{2}$$

Here, χ_c is the *number of colors* in the edge coloring c and $c(e)$ is the *color* of the edge e. Denoting by $\Delta(G)$ the maximum degree of the vertices in G, *Vizing's theorem* proves that $\chi_c \leq \Delta(G) + 1$ for all edge colorings c of G [16].

Now let $G = (V, E)$ be a *directed* graph. That G is directed means E consists of ordered pairs (u, v), which we write as $u \to v$, as opposed to unordered sets $\{u, v\}$. The *in-degree* and *out-degree* of a vertex $v \in V$ are, respectively, the cardinalities of the sets $\{u \to v \in E \mid u \in V\}$ and $\{v \to u \in E \mid u \in V\}$.

A *directed path* in G is a sequence of edges $v_1 \to v_2 \to \cdots \to v_k$ where $k > 1$. We say G is *acyclic* if it does not contain a self-loop $v \to v$ and does not contain a directed path that connects to itself: $v_1 \to v_2 \to \cdots \to v_k \to v_1$. A directed graph that is acyclic is called a *directed acyclic graph* (DAG). Besides an important exception in Sec. 3.4, all graphs in this paper are DAGs.

Given two directed graphs $G = (V, E)$ and $G' = (V', E')$, a *directed graph isomorphism* between G and G' is a bijection $f : V \to V'$ satisfying

$$u \to v \in E \iff f(u) \to f(v) \in E'. \tag{3}$$

Directed graph isomorphisms preserve the structure of directed graphs. Thus, if G and G' are isomorphic as directed graphs, then G is a DAG if and only if G' is. Moreover, they have exactly the same in-degrees/out-degrees, paths, etc. In other words, all that could be different between G and G' are the vertex labels and their meaning [16].

1.2 Probability Theory

Let $\{(\Omega, 2^\Omega, \Pr_1), \ldots, (\Omega, 2^\Omega, \Pr_n)\}$ be a collection of n discrete probability spaces over the alphabet Ω, where 2^Ω is the discrete σ-algebra on Ω (thus all subsets of Ω are measurable but not necessarily supported), and each $\Pr_i : 2^\Omega \to [0,1]$ is a probability measure.

If $(\Sigma, 2^\Sigma)$ is a measurable space over some potentially different alphabet Σ and $\{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$ is a collection of random variables $\mathtt{X}_i : \Omega \to \Sigma$, the *joint random variable* $\mathtt{X} \coloneqq (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ is the map $\mathtt{X} : \Omega^{\times n} \to \Sigma^{\times n}$, where $\Omega^{\times n}$ and $\Sigma^{\times n}$ are n-fold Cartesian products over Ω and Σ, respectively. Formally, any element $x \in \Sigma^{\times n}$ is an n tuple $x = (\sigma_1, \ldots, \sigma_n)$, where each $\sigma_i \in \Sigma$. But of course there is a trivial bijection $(\sigma_1, \ldots, \sigma_n) \mapsto \sigma_1 \ldots \sigma_n \in \Sigma^n \subseteq \Sigma^*$, where Σ^* (Σ^n) is the collection of all (n-length) strings over Σ. Thus, $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ is equivalently a random variable over Σ^n. We therefore make no further distinction between $\Sigma^{\times n}$ and Σ^n (and similarly for $\Omega^{\times n}$ and Ω^n).

Of course, to be a random variable, $\mathtt{X}$ must be a measurable function between measurable spaces, but which measurable spaces? Since each $\mathtt{X}_i$ is a random variable from $(\Omega, 2^\Omega, \Pr_i)$ to $(\Sigma, 2^\Sigma)$, the tuple $(\mathtt{X}_1, \ldots, \mathtt{X}_n)$ naturally induces that $\mathtt{X}$ is a random variable from $(\Omega^n, 2^{\Omega^n}, \Pr)$ to $(\Sigma^n, 2^{\Sigma^n})$, where, generically, the *joint probability measure* $\Pr : \Omega^n \to [0,1]$ is nontrivially related to $\Pr_1, \ldots, \Pr_n$. In particular, Pr is seldom just the product measure $\Pr_1 \times \cdots \times \Pr_n$ because that entails a needlessly strong degree of independence between $\mathtt{X}_1, \ldots, \mathtt{X}_n$.

In practice, Pr is either given or inferred from experimental data, in which case Pr is only ever approximately known. That said, many physical theories posit a network-like structure that dictates which $\mathtt{X}_i$'s can influence which $\mathtt{X}_j$'s (arising, for example, from an underlying causal structure), and this amounts to constraining how Pr can relate to each individual $\Pr_i$. We will return to this idea shortly.

Given a joint probability measure $\Pr : \Omega^n \to [0,1]$ and a joint random variable $\mathtt{X} : \Omega^n \to \Sigma^n$, the *joint distribution over* $\mathtt{X}$, denoted $\Pr_{\mathtt{X} \sim \mathscr{D}}$, is the

pushforward measure of Pr to $(\Sigma^n, 2^{\Sigma^n})$:

$$\forall (S \in 2^{\Sigma^n}) : \Pr_{\mathtt{X} \sim \mathscr{D}}(\mathtt{X} \in S) \coloneqq \Pr(\mathrm{preim}_{\mathtt{X}}(S)). \tag{4}$$

Though ostensibly overdecorated, the notation $\Pr_{\mathtt{X} \sim \mathscr{D}}$ turns out to be linguistically nice because it enables us to refer to a given joint distribution in a myriad of equivalent ways: either via the pushforward measure (4) (which is best for numbered equations) or through any syntactically clear use of the script symbol $\mathscr{D}$ (which is best for referring to Eq. (4) in sentences or for writing certain distributional equations like the total variation distance).

For example, when defining a new joint random variable $\mathtt{X}$, we often say "let $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ be a $\mathscr{D}$-random variable". This is to convey that $\mathtt{X}$ is a joint random variable distributed over $\mathscr{D}$, which just means that $\mathtt{X}$ satisfies Eq. (4). Such notation is coincidentally inspired by terminology like "Bernoulli random" and "Haar random", the notation in complexity theory papers like [17] and [18], and the notation in causal inference papers like [14]. It will also prove useful when equivalent distributions arise out of distinct mathematical objects like probabilistic Turing machines and causal networks.

In this essay, we mostly deal with *valuations* of the joint random variable $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$, which correspond to $\mathtt{X}$ equaling a singular value in Σ^n. For example, $\mathtt{X} = x = \sigma_1 \ldots \sigma_n \in \Sigma^n$ is a valuation (equivalently: $\mathtt{X}_1 = \sigma_1, \ldots, \mathtt{X}_n = \sigma_n$), and we denote its probability as

$$\Pr_{\mathtt{X} \sim \mathscr{D}}(x) \coloneqq \Pr_{\mathtt{X} \sim \mathscr{D}}(\mathtt{X} = x). \tag{5}$$

In probability theory, Eq. (5) is called the *probability mass function* (PMF) of the valuation $\mathtt{X} = x$. Since we only ever deal with valuations of $\mathtt{X}$, we will frequently refer to $\Pr_{\mathtt{X} \sim \mathscr{D}}$ (and other related distributions) as a PMF.

We will frequently deal with valuations that satisfy a particular functional relationship *for all* $x \in \Sigma^n$. If, for example, there is some function f such that $\Pr_{\mathtt{X} \sim \mathscr{D}}(x) = f(x)$ for all $x \in \Sigma^n$, then rather than crowd the notation with universal quantifiers, we will adopt the following rule:

$$\Pr_{\mathtt{X} \sim \mathscr{D}}(\mathtt{X}) = f(\mathtt{X}) \iff \forall (\mathtt{X} = x) : \Pr_{\mathtt{X} \sim \mathscr{D}}(x) = f(x). \tag{6}$$

We illustrate this with the following fact. If

$$\Pr_{\mathtt{X}_i | \mathtt{X}_1, \ldots, \mathtt{X}_{i-1}}(\mathtt{X}_i \mid \mathtt{X}_1, \ldots, \mathtt{X}_{i-1}) \coloneqq \frac{\Pr_{\mathtt{X} \sim \mathscr{D}}(\mathtt{X}_1, \ldots, \mathtt{X}_n)}{\sum_{\mathtt{X}_i = x_i} \cdots \sum_{\mathtt{X}_n = x_n} \Pr_{\mathtt{X} \sim \mathscr{D}}(\mathtt{X}_1, \ldots, \mathtt{X}_{i-1}, x_i, \ldots, x_n)} \tag{7}$$

is the *conditional PMF* of $\mathtt{X}_i$ given $\mathtt{X}_1, \ldots, \mathtt{X}_{i-1}$, then $\Pr_{\mathtt{X}\sim\mathscr{D}}$ satisfies the *probability chain rule* on every valuation [19]:

$$\Pr_{\mathtt{X}\sim\mathscr{D}}(\mathtt{X}_1, \ldots, \mathtt{X}_n) = \prod_i \Pr_{\mathtt{X}_i|\mathtt{X}_1,\ldots,\mathtt{X}_{i-1}}(\mathtt{X}_i \mid \mathtt{X}_1, \ldots, \mathtt{X}_{i-1}). \tag{8}$$

Again, per the rule (6), Eqs. (7) and (8) hold *for all* valuations $\mathtt{X} = x$.

1.3 Bayesian Networks

Whereas Eq. (7) suggests $\mathtt{X}_i$ can depend on any or all of $\mathtt{X}_1, \ldots, \mathtt{X}_{i-1}$, it might happen that variations in some subset $\mathtt{pa}(\mathtt{X}_i) \subseteq \{\mathtt{X}_1, \ldots, \mathtt{X}_{i-1}\}$ necessarily change $\mathtt{X}_i$ (say, in the sense of a direct causal connection) but variations in $\{\mathtt{X}_1, \ldots, \mathtt{X}_{i-1}\}\backslash\mathtt{pa}(\mathtt{X}_i)$ need not (say, in the sense of an indirect causal connection via $\mathtt{pa}(\mathtt{X}_i)$ or no causal connection at all). In this case, Eq. (7) collapses to the simpler equation

$$\Pr_{\mathtt{X}_i|\mathtt{X}_1,\ldots,\mathtt{X}_{i-1}}(\mathtt{X}_i \mid \mathtt{X}_1, \ldots, \mathtt{X}_{i-1}) = \Pr_{\mathtt{X}_i|\mathtt{pa}(\mathtt{X}_i)}(\mathtt{X}_i \mid \mathtt{pa}(\mathtt{X}_i)). \tag{9}$$

If this holds, we say $\mathtt{X}_i$ and $\{\mathtt{X}_1, \ldots, \mathtt{X}_{i-1}\}\backslash\mathtt{pa}(\mathtt{X}_i)$ are *conditionally independent given* $\mathtt{pa}(\mathtt{X}_i)$, which we alternatively express as

$$\mathtt{X}_i \perp\!\!\!\perp \{\mathtt{X}_1, \ldots, \mathtt{X}_{i-1}\}\backslash\mathtt{pa}(\mathtt{X}_i) \mid \mathtt{pa}(\mathtt{X}_i). \tag{10}$$

We call the set $\mathtt{pa}(\mathtt{X}_i)$ the *Markovian parents* of $\mathtt{X}_i$.

In this essay we impose the much stronger *directed local Markov property*:

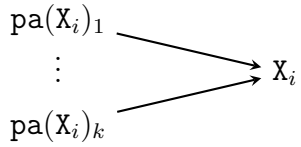
$$\mathtt{X}_i \perp\!\!\!\perp \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}\backslash\mathtt{nd}(\mathtt{X}_i) \mid \mathtt{pa}(\mathtt{X}_i), \tag{11}$$

where $\mathtt{nd}(\mathtt{X}_i) \subseteq \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$ is the set of *Markovian non-descendants* of $\mathtt{X}_i$. Intuitively, $\mathtt{nd}(\mathtt{X}_i)$ consists of all variables that could depend on $\mathtt{X}_i$. Physically, the directed local Markov property is justified by the intuition that $\mathtt{X}_i$ ought only to directly depend on the variables $\mathtt{pa}(\mathtt{X}_i)$ in its immediate past and at most indirectly on any variables $\{\mathtt{X}_1, \ldots, \mathtt{X}_n\}\backslash\mathtt{nd}(\mathtt{X}_i)$ not in its future, in its far past, or outside its past altogether.

Thinking causally, Eq. (11) admits the intuitive graphical representation:

$$\mathtt{pa}(\mathtt{X}_i) \longrightarrow \mathtt{X}_i \tag{12}$$

which by construction encodes the conditional independence (CI) relations of $\mathtt{X}_i$. Of course, we can enlarge (12) by aiming each $\mathtt{pa}(\mathtt{X}_i)_j \in \mathtt{pa}(\mathtt{X}_i)$ toward $\mathtt{X}_i$. Then, if $\mathtt{pa}(\mathtt{X}_i)$ has k elements $\mathtt{pa}(\mathtt{X}_i)_1, \ldots, \mathtt{pa}(\mathtt{X}_i)_k$, (12) becomes:



(13)

This graph is for a single $\mathtt{X}_i \in \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$. But of course there is no impediment in the way of generalizing (13) to each $\mathtt{X}_i \in \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$ (provided we assume the directed local Markov property for each, which we do). To this end, Eq. (9) collapses Eq. (8) to the simpler equation

$$\Pr_{\mathtt{X}\sim\mathscr{D}}(\mathtt{X}_1, \ldots, \mathtt{X}_n) = \prod_i \Pr_{\mathtt{X}_i|\mathtt{pa}(\mathtt{X}_i)}(\mathtt{X}_i \mid \mathtt{pa}(\mathtt{X}_i)), \tag{14}$$

where now each $\mathtt{X}_i$ only depends on its Markovian parents $\mathtt{pa}(\mathtt{X}_i)$. Equivalently, Eq. (14) corresponds to the DAG $G = (V, E)$, where the vertex set $V = \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$ and the edge set E satisfies

$$\mathtt{X}_j \longrightarrow \mathtt{X}_i \in E \iff \mathtt{X}_j \in \mathtt{pa}(\mathtt{X}_i). \tag{15}$$

If a DAG G represent the CI relations of a distribution $\mathscr{D}$, then G is called *Bayesian network for* $\mathscr{D}$.

Definition 1 (Bayesian Network)**.** Let $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ be a $\mathscr{D}$-random variable and $G = (V, E)$ a DAG. We say G is a *Bayesian network for* $\mathscr{D}$ or a *DAG for* $\mathscr{D}$ if and only if $\Pr_{\mathtt{X}\sim\mathscr{D}}$ factorizes in the form of (16) and there exists a directed graph isomorphism between G and the DAG $\widetilde{G} = (\widetilde{V}, \widetilde{E})$, where $\widetilde{V} = \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$ and $\widetilde{E}$ satisfies Eq. (15).

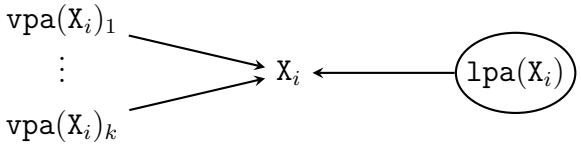
1.4 Marginal Bayesian Networks

There is an important generalization of Bayesian networks that is captured by the following fact: any PMF satisfying Eq. (14) also satisfies

$$\Pr_{\mathtt{X}\sim\mathscr{D}}(\mathtt{X}) = \sum_{\Lambda_1=\lambda_1} \cdots \sum_{\Lambda_m=\lambda_m} \prod_{i,j} \Pr_{\mathtt{X}_i|\mathtt{pa}(\mathtt{X}_i),\Lambda_1,\ldots,\Lambda_m}(\mathtt{X}_i \mid \mathtt{pa}(\mathtt{X}_i), \lambda_1, \ldots, \lambda_m) \Pr_{\Lambda_j\sim\mathscr{R}_j}(\lambda_j), \tag{16}$$

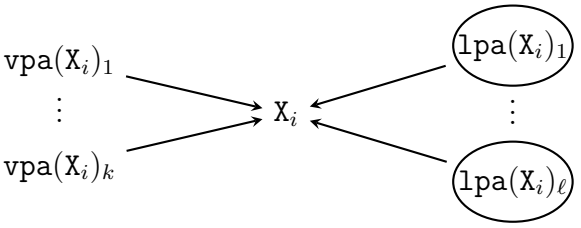
where each Λ_j is a discrete $\mathscr{R}_j$-random *latent* or *hidden* variable, in the sense that it is marginalized out in the PMF (16). This statement is trivial, because to recover Eq. (14) one need only impose that each $\mathscr{R}_j$ be a point distribution. That said, the converse is not generally true: a PMF satisfying Eq. (16) need not satisfy Eq. (14).

Let $L = \{\Lambda_1, \ldots, \Lambda_m\}$ be the set of latent variables, and define $\mathtt{lpa}(\mathtt{X}_i) = L \cap \mathtt{pa}(\mathtt{X}_i)$ as the set of *latent Markovian parents* of $\mathtt{X}_i$. Then, with the addition of the latent variables L and supposing, as always, that all $\mathtt{X}_i \in \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$ and $\Lambda_j \in L$ obey the directed local Markov property, there are now two possibilities. Either $\mathtt{lpa}(\mathtt{X}_i) = \varnothing$, in which case $\mathtt{X}_i$ has no latent parents and so $\mathtt{X}_i \perp\!\!\!\perp \{\mathtt{X}_1, \ldots, \mathtt{X}_n\} \cup L \backslash \mathtt{nd}(\mathtt{X}_i) \mid \mathtt{pa}(\mathtt{X}_i)$ implies (13) as before, or $\mathtt{lpa}(\mathtt{X}_i) \neq \varnothing$, in which case $\mathtt{X}_i$ has latent parents and so $\mathtt{X}_i \perp\!\!\!\perp \{\mathtt{X}_1, \ldots, \mathtt{X}_n\} \cup L \backslash \mathtt{nd}(\mathtt{X}_i) \mid \mathtt{pa}(\mathtt{X}_i)$ does not imply (13). In this latter case, the correct graph is instead,



(17)

where the ellipse signifies latency and $\mathtt{vpa}(\mathtt{X}_i) = \{\mathtt{X}_1, \ldots, \mathtt{X}_n\} \cap \mathtt{pa}(\mathtt{X}_i)$ is the set of *visible* (i.e., non-latent) Markovian parents of $\mathtt{X}_i$. As before, we can enlarge (17) by aiming each $\mathtt{lpa}(\mathtt{X}_i)_1, \ldots, \mathtt{lpa}(\mathtt{X}_i)_\ell \in \mathtt{lpa}(\mathtt{X}_i)$ toward $\mathtt{X}_i$:



(18)

Proceeding as before for each $\mathtt{X}_i \in \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$, one obtains an equivalent characterization of Eq. (16) as a DAG $G = (V \cup L, E)$, where the *visible vertices* $V = \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$, the *latent vertices* $L = \{\Lambda_1, \ldots, \Lambda_n\}$, and the edge set E satisfies

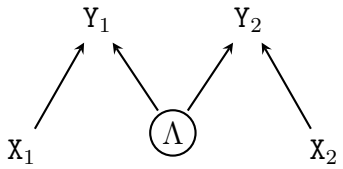
$$\begin{cases} \mathtt{X}_j \longrightarrow \mathtt{X}_i \in E \iff \mathtt{X}_j \in \mathtt{vpa}(\mathtt{X}_i), \\ \textcircled{\Lambda_j} \longrightarrow \mathtt{X}_i \in E \iff \Lambda_j \in \mathtt{lpa}(\mathtt{X}_i). \end{cases} \tag{19}$$

Following Evans [20, 21], any Bayesian network that represents a marginalized distribution like Eq. (16) is called a *marginalized Bayesian network*, *marginalized DAG*, or *mDAG* for short.

Definition 2 (Marginal Bayesian Network)**.** Let $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ be a $\mathscr{D}$-random variable and $G = (V, E)$ a DAG. We say G is a *marginal Bayesian network for* $\mathscr{D}$ or an *mDAG for* $\mathscr{D}$ if and only if $\Pr_{\mathtt{X} \sim \mathscr{D}}$ factorizes in the form of (16) and there exists a directed graph isomorphism between G and the mDAG $\widetilde{G} = (\widetilde{V} \cup L, \widetilde{E})$, where $\widetilde{V} = \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$, $L = \{\Lambda_1, \ldots, \Lambda_m\}$, and $\widetilde{E}$ satisfies Eq. (19).

Importantly, if G is an mDAG for a distribution $\mathscr{D}$, then $\mathscr{D}$ is not in any reasonable sense "unique" to G. That is, there always exists a distribution $\mathscr{D}' \neq \mathscr{D}$ for which G is an mDAG. This must be true, because CI relations are impartial to the cardinalities of the distributions they constrain. That said, there also exist distributions for which G is *not* an mDAG. This is true because different distributions can have different CI relations. There is thus a natural sense in which a distribution $\mathscr{D}$ is *compatible* with a given mDAG G. In particular, if $\mathscr{P}(G)$ denotes the set of all distributions $\mathscr{D}$ for which G is an mDAG, then we say $\mathscr{D}$ is *compatible* with G if and only if $\mathscr{D} \in \mathscr{P}(G)$.

To give an example, consider the mDAG known as the *Bell scenario*:



(20)

Thinking causally (though such thinking is hitherto unjustified), the Bell scenario encodes the expected causal relations for any experiment in which the pairs $\mathtt{Y}_1$ and $\mathtt{Y}_2$ are spacelike separated and only have Λ as a (hidden) common ancestor (in particular, $\mathtt{X}_1$ and $\mathtt{X}_2$ have nothing to do with each other, and are also spacelike separated). Consequently, the Bell scenario encodes the physically-motivated CI relations:

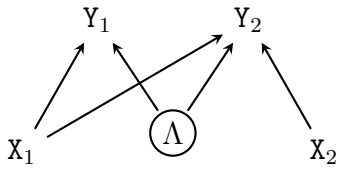
$$\mathtt{X}_1 \perp\!\!\!\perp \mathtt{X}_2, \Lambda \qquad \text{and} \qquad \mathtt{X}_2 \perp\!\!\!\perp \mathtt{X}_1, \Lambda, \tag{21}$$

$$\mathtt{Y}_1 \perp\!\!\!\perp \mathtt{Y}_2, \mathtt{X}_2 \mid \mathtt{X}_1, \Lambda \quad \text{and} \quad \mathtt{Y}_2 \perp\!\!\!\perp \mathtt{Y}_1, \mathtt{X}_1 \mid \mathtt{X}_2, \Lambda, \tag{22}$$

which, respectively, are called the *no superdeterminism constraint*, which can be construed as a statement about an experimenter's free will [14], and the *local causality constraint*, which entails that $\mathtt{Y}_1$ ought only to depend on events in its backward lightcone (and similarly for $\mathtt{Y}_2$).

Central to this physical interpretation, however, is that the set of distributions compatible with the Bell scenario is a *strict* subset of the distributions

compatible with the *Bell scenario with communication*:



(23)

Within (23), there is now potential for a direct causal influence to propagate from $\mathtt{X}_1$ to $\mathtt{Y}_2$, which is formally forbidden in the standard Bell scenario (20). The most direct way to see this is to marginalize $\mathtt{Y}_2$ and Λ in the first CI relation in (22), and $\mathtt{Y}_1$ and Λ in the second CI relation in (22). This reveals that the Bell scenario also encodes the CI relations

$$\mathtt{Y}_1 \perp\!\!\!\perp \mathtt{X}_2 \mid \mathtt{X}_1 \quad \text{and} \quad \mathtt{Y}_2 \perp\!\!\!\perp \mathtt{X}_1 \mid \mathtt{X}_2. \tag{24}$$

Physically speaking, these correspond to the *no superluminal signalling* constraint [14]. As Eq. (24) is manifestly violated in the Bell scenario with communication, it is plain that the Bell scenario with communication has strictly more compatible distributions than the Bell scenario. Incidentally, this example also illustrates the more general fact that a given mDAG often entails more CI relations than meet the eye. While there is a purely graphical way to "see" all possible CI relations of a given mDAG (it is called *d-separation* [19]), it is incidental for the purposes of this essay.

1.5 Semialgebraic Statistics

In this section, we bridge a connection between statistics and algebraic geometry—a connection that is formally pursued in the theory of *algebraic statistics* [22, 23]. For us, the tools of algebraic statistics afford a complete *algebraic* characterization of CI relations in terms of equality and inequality constraints. At a high level, this implies a rather profound equivalence between structures in graph theory, statistics, and algebraic geometry. Incidentally, as we explore in Sec. 3, it is precisely this three-way relationship that connects Bayesian networks to shallow circuits.

Let $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ be a $\mathscr{D}$-random variable over the alphabet Σ. There are then $D = |\Sigma|^n$ unique valuations of $\mathtt{X}$, where the ith valuation $\mathtt{X} = x_i$ has probability $p_i \coloneqq \Pr_{\mathtt{X} \sim \mathscr{D}}(x_i)$. Collectively, the tuple $\boldsymbol{p} = (p_1, \ldots, p_D)$ is a

point in the *probability simplex*,

$$\Delta_{\Sigma^n} \coloneqq \left\{ \boldsymbol{p} \in \mathbb{R}^D \;\middle|\; p_i \geq 0,\ \sum_{i=1}^{D} p_i = 1 \right\}. \tag{25}$$

Now suppose $\mathtt{X}$ is distributed differently over Σ, which is to say its $\mathscr{D}'$-random for some $\mathscr{D}' \neq \mathscr{D}$. In general, then, the associated point $\boldsymbol{p}' = (p'_1, \ldots, p'_D)$ is different from $\boldsymbol{p}$, and hence $\boldsymbol{p}'$ lands elsewhere in Δ_{Σ^n}. Imagining all the ways in which $\mathtt{X}$ can be distributed over Σ ($\mathscr{D}''$, $\mathscr{D}'''$, and so on) produces a set $\mathcal{M} \coloneqq \{\boldsymbol{p}, \boldsymbol{p}', \boldsymbol{p}'', \boldsymbol{p}''', \ldots\}$ that is a subset of the probability simplex Δ_{Σ^n}. We call $\mathcal{M}$ a *discrete statistical model* of $\mathtt{X}$.

If G is an mDAG with compatible distributions $\mathscr{P}(G)$, then, per the argument above, there exists a statistical model $\mathcal{M}(G) \subseteq \Delta_{\Sigma^n}$ that characterizes $\mathscr{P}(G)$ geometrically. Thus, to every mDAG G there corresponds a discrete statistical model $\mathcal{M}(G)$. In particular, the d-separation properties of G, and hence the CI relations embedded in G, must in some way be geometrically encoded in $\mathcal{M}(G)$.

To understand how, consider first the CI relation $\mathtt{X}_i \perp\!\!\!\perp \mathtt{X}_j \mid \mathtt{X}_k$ for some $i \neq j \neq k$. Then,

$$\Pr_{\mathtt{X}_i|\mathtt{X}_k,\mathtt{X}_j}(\mathtt{X}_i \mid \mathtt{X}_k, \mathtt{X}_j) = \Pr_{\mathtt{X}_i|\mathtt{X}_k}(\mathtt{X}_i \mid \mathtt{X}_k). \tag{26}$$

The definition of the conditional PMF (7) implies Eq. (26) is equivalent to

$$\Pr_{\mathtt{X}_i,\mathtt{X}_k,\mathtt{X}_j}(\mathtt{X}_i, \mathtt{X}_k, \mathtt{X}_j)\Pr_{\mathtt{X}_k}(\mathtt{X}_k) - \Pr_{\mathtt{X}_i,\mathtt{X}_k}(\mathtt{X}_i, \mathtt{X}_k)\Pr_{\mathtt{X}_k,\mathtt{X}_j}(\mathtt{X}_k, \mathtt{X}_j) = 0. \tag{27}$$

Thus, the CI relation $\mathtt{X}_i \perp\!\!\!\perp \mathtt{X}_j \mid \mathtt{X}_k$ is equivalent to a polynomial equality constraint over the marginal distributions $\Pr_{\mathtt{X}_i,\mathtt{X}_k,\mathtt{X}_j}$, $\Pr_{\mathtt{X}_k,\mathtt{X}_j}$, $\Pr_{\mathtt{X}_i,\mathtt{X}_k}$, and $\Pr_{\mathtt{X}_k}$. Geometrically, if $\mathtt{X}_i \perp\!\!\!\perp \mathtt{X}_j \mid \mathtt{X}_k$ holds for all distributions over $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$, then it will be geometrically encoded into the discrete statistical model $\mathcal{M}$ of $\mathtt{X}$ by restricting $\mathcal{M}$ to the hypersurface defined by Eq. (27).

Interestingly, that CI relations are polynomial equality constraints is a general phenomenon. In particular, if $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ is a $\mathscr{D}$-random variable and $\mathtt{X}_D \coloneqq (\mathtt{X}_{d_1}, \ldots, \mathtt{X}_{d_k})$ is the joint random variable formed by any subset $D = \{\mathtt{X}_{d_1}, \ldots, \mathtt{X}_{d_k}\} \subseteq \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$, then it can be shown that $\Pr_{\mathtt{X} \sim \mathscr{D}}$ satisfies $A \perp\!\!\!\perp B \mid C$ for disjoint sets $A, B, C \subsetneq \{\mathtt{X}_1, \ldots, \mathtt{X}_n\}$ if and only if

$$\begin{aligned}\Pr_{A,B,C}(x_A, x_B, x_C)\Pr_{A,B,C}(x'_A, x'_B, x_C) \\ - \Pr_{A,B,C}(x'_A, x_B, x_C)\Pr_{A,B,C}(x_A, x'_B, x_C) = 0\end{aligned} \tag{28}$$

holds for every valuation $\mathtt{X}_C = x_C$, $\mathtt{X}_B = x_B, x'_B$, and $\mathtt{X}_A = x_A, x'_A$ [23, 24]. Consequently, if, for some random variable $\mathtt{X}$, a general CI relation of the form $A \perp\!\!\!\perp B \mid C$ holds for *every* distribution over $\mathtt{X}$, then the discrete statistical model $\mathcal{M}$ of $\mathtt{X}$ will intersect the hypersurface defined by Eq. (28).

Now consider a latent-free Bayesian network $G = (V, E)$ and denote by $\mathbb{R}[p_1, \ldots, p_D]$ the set of all polynomials in indeterminates $p_1, \ldots, p_D$ (that is, the *polynomial ring* over the field $\mathbb{R}$). If $\mathrm{CI}(G)$ is the set of all CI relations entailed by G (derived, say, by the d-separation criterion), then, by Eq. (28), to each CI relation $(A_i \perp\!\!\!\perp B_i \mid C_i) \in \mathrm{CI}(G)$ there corresponds a polynomial $P_i \in \mathbb{R}[p_1, \ldots, p_D]$ such that $P_i(\boldsymbol{p}) = 0$ if and only if $\boldsymbol{p} \in \mathcal{M}(G)$. In other words, $\mathcal{M}(G)$ is the *algebraic variety* defined by the set $\{P_1, \ldots, P_{|\mathrm{CI}(G)|}\} \subseteq \mathbb{R}[p_1, \ldots, p_D]$. Therefore, the statistical model $\mathcal{M}(G)$, and hence the distributions compatible with G, is equivalent to the vanishing set of the polynomials $P_1, \ldots, P_{|\mathrm{CI}(G)|}$. We can therefore reason about $\mathscr{P}(G)$ by reasoning about these polynomials [22, 25]. In particular, for mDAGs G and G', it holds that $\mathscr{P}(G) \subseteq \mathscr{P}(G')$ if and only if $\mathcal{M}(G) \subseteq \mathcal{M}(G')$.

Of course, the preceding analysis only applies in the case of a latent-free Bayesian network. Fortunately, the extension to marginal Bayesian networks is not too hard. The main insight is that there is a natural directed graph isomorphism from any marginal Bayesian network to a latent-free Bayesian network: simply remove the circles from the mDAG. Thus, given any mDAG $G = (V \cup L, E)$, there corresponds an isomorphic DAG $\widetilde{G} = (\widetilde{V}, \widetilde{E})$. Importantly, $\widetilde{G}$ is also a valid Bayesian network with a PMF equal to Eq. (16) but without the marginalization over the latent variables (thus to obtain the PMF for G from $\widetilde{G}$, simply marginalize the latent variables).

For simplicity, let us suppose there are m latents $\Lambda_1, \ldots, \Lambda_m \in L$ and that every Λ_j is a random variable over the same alphabet Σ. Further, put $D' = |\Sigma^m|$. Then the CI relations of the Bayesian network $\widetilde{G}$ correspond to a set of polynomials $\mathcal{F}$ in indeterminates $p_1, \ldots, p_{D+D'}$. That is, $\mathcal{F} \subseteq \mathbb{R}[p_1, \ldots, p_{D+D'}]$. As before, $\mathcal{M}(\widetilde{G})$ is the algebraic variety over this set,

$$\mathcal{M}(\widetilde{G}) = \left\{\boldsymbol{p} \in \Delta_{\Sigma^{n+m}} \mid P(\boldsymbol{p}) = 0, P \in \mathcal{F}\right\}. \tag{29}$$

We have thus derived the discrete statistical model for the Bayesian network $\widetilde{G}$. Evidently, $\mathcal{M}(\widetilde{G})$ is a higher dimensional hypersurface than $\mathcal{M}(G)$, and yet they both obey the exact same CI relations. The additional dimensions, therefore, owe to treating the latent variables as visible vertices. The restriction to $\mathcal{M}(G)$ is done by projecting $\mathcal{M}(\widetilde{G})$ into

the hypersurface defined by $\mathcal{M}(G)$. That is, we define a projection map $\pi : (p_1, \ldots, p_{D+D'}) \mapsto (p_1, \ldots, p_D)$ so that $\pi(\mathcal{M}(\widetilde{G})) = \mathcal{M}(G)$ [21, 23, 26].

Following a projection like this, it is not obvious that $\mathcal{M}(G)$ will correspond to an algebraic variety of some set of polynomials. In fact, this is hardly ever the case [26].[1] Instead, according to the highly nontrivial *Tarski-Seidenberg theorem* [26], the projection of an algebraic variety into a subset of its coordinates is at best a *semialgebraic set*, which is a finite collection of polynomial equalities and polynomial inequalities. The following theorem is thus a simple corollary of the Tarski-Seidenberg theorem:

Theorem 1. *If G is an mDAG, then $\mathcal{M}(G)$ is a semialgebraic set.*

In other words, for any mDAG G, its statistical model $\mathcal{M}(G)$, and hence its compatible distributions $\mathscr{P}(G)$, correspond to a set of points $\boldsymbol{p} \in \Delta_{\Sigma^n}$ that satisfy a finite set of polynomial equality and inequality constraints. This idea underlies everything to come.

[1] Consider, for example, projecting the algebraic variety $\{(x, y) \in \mathbb{R}^2 \mid xy - 1 = 0\}$ into the x axis. The result is $\{x \in \mathbb{R} \mid |x| > 0\}$, which is evidently not an algebraic variety. (It is, however, a semialgebraic set.)

2 Clarifying "Quantum" with Causal Inference

In this section, we define "quantum" in a way that entails a violation of a Bell-type inequality. We frame this discussion using tools from causal inference that are inextricably tied to the theory of Bayesian networks. We close with a discussion on "quantum-classical gaps" and give two pertinent examples.

2.1 Classical and Quantum Operational Theories

Contrary to most textbooks, a theory is hardly "quantum" if it entails interference, superposition, entanglement, no-cloning, teleportation, or other *canonically quantum* phenomena. This follows from the existence of decidedly classical theories like [27] and [28] that exhibit most if not all canonically quantum phenomena. There is thus a nontrivial demarcation problem to address: which physical theories are "quantum" and which are "classical"?[2]

Clearly, any serious demarcation criterion must presuppose some agreeable definition of "physical theory". One primitive definition is as a body of sentences, expressed as a formal language over some alphabet of symbols. This so-called *syntactic view of theories* was popularized in the 20th century by logical positivists like Carnap and Hempel [29]. Whereas the sentiment that this is *all* a physical theory is is predictably controversial, the sentiment that this is *at least* what a physical theory is is arguably less so.

In addition to a language, one wants a semantics, or an interpretation of the language, and moreover one wants that interpretation to comport with one's experience in the laboratory. Given a syntactic theory, one minimal interpretation is as a *prepare-and-measure operational theory*. That is, the syntax of the theory relates to preparation and measurement procedures ("laboratory instructions") and also provides a rule for predicting the frequencies of different measurement outcomes (e.g., the Born rule). Thus, a prepare-and-measure operational theory is a mathematical framework for predicting the outcomes of a prepare-and-measure experimental procedure [30].

In a *classical theory*, the rule for predicting frequencies uses the standard probability calculus. A *quantum theory*, however, uses Hilbert spaces, density operators, positive operator-valued measures, and the like. At the end of the day, however, both classical and quantum theories give rise to

[2]We are omitting an obvious and interesting third category: "neither". Fortunately, the forthcoming techniques work just as well for the full, three-way demarcation problem [12].

a conditional probability measure $\Pr(m_k \mid P, M)$ that equals the likelihood of obtaining one of a discrete set of measurement outcomes $\{m_k\}$ given a particular preparation P and measurement M.

If, for a given preparation and measurement procedure P and M, a distribution of measurement outcomes $\mathscr{D} = \{\Pr(m_k \mid P, M)\}$ is all one has, then distinguishing a classical and a quantum theory is impossible. For example, one cannot distinguish the distribution entailed by the flip of a fair coin and the distribution stemming from measurements of the state $\frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$ in the computational basis. However, given a joint distribution of measurements $\mathscr{D} = \{\Pr(m_{k_1}, m_{k_2}, \dots \mid P_1, M_1, P_2, M_2, \dots)\}$, it becomes possible to analyze correlations given conditional independence assumptions. In particular, the correlations allowed quantumly may transcend those that are allowed classically. If such correlations exist, then they would constitute a *quintessentially quantum* phenomenon because no classical theory can reproduce them.

As correlations were central to the theories in the introductory parts of this essay, it is no surprise that a useful way to examine prepare-and-measure operational theories is with Bayesian networks and semialgebraic statistics.

2.2 Causal Networks and Causal Compatibility

As was hinted at in Sec. 1.3, it is fine for our purposes to interpret a marginal Bayesian network $G = (V \cup L, E)$ as not merely a graphical encoding of the CI relations among the vertices $V \cup L$, but also as an encoding of the *causal relations* among $V \cup L$. Strictly speaking, this interpretation is not justified because CI relations are *associational* relations and these are only rarely causal in nature [19]. For instance, to say temperature and ice cream sales are correlated is one thing, but to say temperature *causes* ice cream sales is another matter entirely. This point is none other than the oft-quoted slogan: correlation does not imply causation.

That said, all the random variables in this essay have origins in an operationally-defined physical theory, which we will take to presuppose a causal structure. This is justified, at least in part, because causal relationships are sensibly more "stable" than any sort of non-causal, associational relationship like temperature and ice cream.[3] (Indeed, it was hot in June 2021, but ice cream shops were closed because of COVID-19.) Thus, insofar as one expects a physical theory to make stable probabilisitic predictions, it is

[3]This is related to the *do-calculus* of Pearl [19]

natural to suppose that the statistical correlations stemming from the theory arise causally. In this way any random variables rooted in an operationally-defined physical theory will satisfy the directed local Markov property, which, when speaking causally, is often called the *causal Markov condition* [13].

Much of the motivation for this more causal interpretation is based on *Reichenbach's common cause principle*, which stipulates that while correlation does not imply causation (in the sense that $\mathtt{X}_1$ correlated with $\mathtt{X}_2$ does not imply $\mathtt{X}_1$ causally entails $\mathtt{X}_2$), correlation does imply a cause-effect relation or a common cause (in the sense that $\mathtt{X}_1$ correlated with $\mathtt{X}_2$ implies either that there exists a common element in $\mathtt{pa}(\mathtt{X}_1)$ and $\mathtt{pa}(\mathtt{X}_2)$, in which case this common element is a candidate for the common cause, or there exists an element in $\mathtt{pa}(\mathtt{X}_1)$ and an element in $\mathtt{pa}(\mathtt{X}_2)$ that are correlated, in which case we recursively consider their parents and again apply Reichenbach's principle).

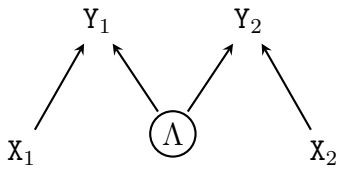
Of course, Reichenbach's principle is just that, a principle. It does not define "causal". Its virtue, rather, is in pinpointing an indispensable property that any agreeable definition of "causal" ought to have—namely, that correlations can be explained causally. Reichenbach's principle, therefore, is what undergirds the idea that causal relations obey the causal Markov condition, and so it is, in a sense, the philosophical principle that justifies the use of DAGs to talk about correlations in a causal way [13].

As Reichenbach's principle is a metaphysical principle, how it manifests in a particular physical theory is a function of that theory. This is because what qualifies as a "common cause" to observed correlations is ultimately a function of the theory.

In a classical theory, for instance, a common cause is a latent variable, exactly like in the mDAGs of Sec. 1.4. So, given some joint distribution of measurement statistics $\mathscr{D}$, we can infer if that distribution arose causally out of a classical theory using an mDAG structure. Formally, we say an mDAG $G = (V \cup L, E)$ is a *classical causal network* for $\mathscr{D}$ if and only if the CI relations entailed by G reflect the causal relations among the vertices $V \cup L$ and $\mathscr{D} \in \mathscr{P}(G)$.

In a quantum theory, however, a common cause is either a (classical) latent variable or a quantum system. Thus, the causal networks in a quantum theory are decidedly different from those in a classical causal network. Rather than attempt the intricate task of defining a *quantum causal network* [31, 32], it suffices for this essay to appeal to the idea that to each classical causal network there corresponds a natural quantum network in which a subset of the latent variables "go quantum".

As an example, recall the Bell scenario from Sec. 1.4:



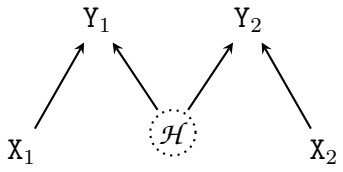
(30)

This is a classical causal network that entails the following conditional distribution

$$\Pr_{\mathtt{Y}|\mathtt{X}\sim\mathscr{D}(\text{Bell})}(\mathtt{Y}_1,\mathtt{Y}_2 \mid \mathtt{X}_1,\mathtt{X}_2) \coloneqq \frac{\Pr_{\mathtt{Y},\mathtt{X}}(\mathtt{Y}_1,\mathtt{Y}_2,\mathtt{X}_1,\mathtt{X}_2)}{\sum_{\mathtt{X}_1=x_1}\sum_{\mathtt{X}_2=x_2}\Pr_{\mathtt{Y},\mathtt{X}}(\mathtt{Y}_1,\mathtt{Y}_2,x_1,x_2)},$$

where $\Pr_{\mathtt{X},\mathtt{Y}}$ is the joint distribution (16) implied by the Bell mDAG (30).

However, there is a natural qDAG associated to this situation (which we call *QBell*) that is obtained by replacing the latent variable Λ (a classical source) with a bipartite Hilbert space $\mathcal{H} = \mathcal{H}_1 \otimes \mathcal{H}_2$ (a quantum source):



(31)

Here, the dotted circle indicates that the source is quantum. Quantum mechanically, then, QBell entails the conditional probability distribution

$$\Pr_{\mathtt{Y}|\mathtt{X}\sim\mathscr{D}(\text{QBell})}(\mathtt{Y}_1,\mathtt{Y}_2 \mid \mathtt{X}_1,\mathtt{X}_2) = \operatorname{Tr}\big(\rho_{12} M_{\mathtt{Y}_1|\mathtt{X}_1} \otimes M_{\mathtt{Y}_2|\mathtt{X}_2}\big), \tag{32}$$

where $M_{\mathtt{Y}_1|\mathtt{X}_1}$ and $M_{\mathtt{Y}_2|\mathtt{X}_2}$ are measurement operators that depend on the random variables $\mathtt{X}_1$ and $\mathtt{X}_2$, respectively. Like in the case of mDAGs, there corresponds a set of distributions $\mathscr{P}(\text{QBell})$ that are compatible with QBell.

This idea generalizes in the obvious way. Given an mDAG G (interpreted as a classical causal network), one can contrive a qDAG QG by making one of the classical latent sources quantum mechanical. Of course, one can be even more general and compare an mDAG G to an unrelated qDAG QG'.

2.3 QC-Gaps as Quintessentially Quantum Phenomena

A natural question is if every distribution compatible with a qDAG is compatible with the mDAG on which it is based. For example, is QBell compatible with Bell? That is, does $\mathscr{P}(\text{Bell}) = \mathscr{P}(\text{QBell})$? If this is true, then, at least in the Bell scenario, there is no discernible difference between the measurement statistics entailed by classical theory and quantum theory because all the correlations are the same. However, if $\mathscr{P}(\text{Bell}) \neq \mathscr{P}(\text{QBell})$, then classical and quantum theory make disparate predictions about the nature of correlations over spacelike distances. In consequence, our classical intuitions would be challenged like never before and lie embedded in a concrete of confusion for decades forth.

As such confusion stokes us all, the claim in question is familiarly false. This is the content of John Bell's eponymous theorem [33, 34]:

Theorem 2 (Bell's Theorem)**.** $\mathscr{P}(\text{Bell}) \subsetneq \mathscr{P}(\text{QBell})$.

In other words, the distributions compatible with the quantum formalism in the Bell scenario are strictly more than the distributions compatible with the classical formalism. Bell's theorem thus witnesses a quintessentially quantum phenomenon—a phenomenon that is "quantum-complete" unlike the phenomena of superposition, entanglement, and so forth.

More generally, Bell's theorem is an instance of a *quantum-classical gap*:

Definition 3 (Quantum-Classical Gap)**.** An mDAG G admits a *quantum-classical gap* (QC-gap) if and only if $\mathscr{P}(G) \subsetneq \mathscr{P}(QG)$.

Generically, it is hard to figure if an mDAG G admits a QC-gap or not. That said, there is an elementary sufficient condition that is quite useful:

Lemma 3. *Let G be an mDAG with statistical model $\mathcal{M}(G)$ and suppose $\mathcal{B} \leq 0$ is an inequality in the semialgebraic set that characterizes $\mathcal{M}(G)$ (that is, for all $\boldsymbol{p} \in \mathcal{M}(G)$ it holds that $\mathcal{B}(\boldsymbol{p}) \leq 0$). If there exists $\boldsymbol{p}' \in \mathcal{M}(QG)$ such that $\mathcal{B}(\boldsymbol{p}') > 0$, then G admits a QC-gap.*

The proof is trivial.

Despite being couched in new terminology, Lemma 3 is hardly a new idea: Clauser, Horne, Shimony, and Holt (CHSH) used precisely this logic to prove Bell's theorem in order to make it more amenable to an experimental test [33, 35]. The following definition is thus historically motivated:

Definition 4 (Generalized Bell Inequality)**.** If $\mathcal{B}$ is a polynomial inequality constraint of an mDAG G and if the premises of Lemma 3 hold, then $\mathcal{B}$ is a *generalized Bell inequality.*[4]

Whereas Lemma 3 is somewhat obvious, its converse is not. Surprisingly, though, a result of Henson, Lal, and Pusey (HLP) together with a theorem of Evans proves it true [12, 20]:

Theorem 4. *An mDAG G admits a QC-gap if and only if $\mathcal{M}(G)$ is constrained by a generalized Bell inequality $\mathcal{B}$.*

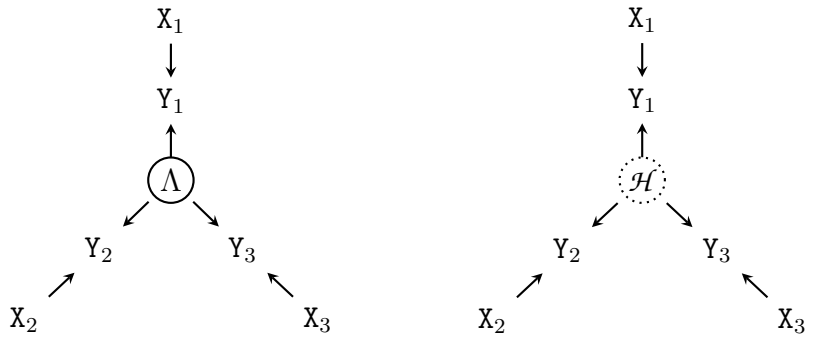
In other words, the equality constraints in the semialgebraic set characterizing $\mathcal{M}(G)$ are exactly satisfied by the statistical model $\mathcal{M}(QG)$. Consequently, the existence of a generalized Bell inequality is a sufficient and necessary condition for a QC-gap.

Unfortunately, however, actually finding a generalized Bell inequality is computationally hard, and presently only a handful of mDAGs are known to support them [12, 15, 37]. Thus, while finding a generalized Bell inequality is equivalent to proving a QC-gap, this does not make proving a QC-gap easy.

That said, an enticing idea to prove a QC-gap is to bootstrap off of mDAGs that already admit QC-gaps (e.g., Bell). More precisely, given mDAGs G and $\check{G}$ related by a "QC-gap non-increasing" map from G to $\check{G}$, it suffices to prove $\check{G}$ admits a QC-gap to prove G admits a QC-gap. This is exactly the idea forthcoming in [15] (see Appendix A).

2.4 Two Examples: GHZ_n and $\mathrm{GHZ}_{n,m}$

Consider the mDAG GHZ_3 and its quantum generalization $QGHZ_3$:



(33)

[4] We acknowledge that the authors of [36] might quibble with this terminology.

This is a generalization of the Bell scenario to three parties and corresponds to the famous 3-party Greenberger-Horne-Zeilinger (GHZ) experiment.[5]

Like in Bell, one can imagine $\mathtt{Y}_1, \mathtt{Y}_2$, and $\mathtt{Y}_3$ as spacelike separated observers, with $\mathtt{X}_1, \mathtt{X}_2$, and $\mathtt{X}_3$ variables dictating the measurement they make on the latent classical or quantum source. Famously, GHZ_3 admits a QC-gap [38].

The generalization from three parties to n parties is the next natural step to take. Using techniques from [15], it is easy to bootstrap off of the QC-gap of Bell to prove that GHZ_n must also admit a QC-gap (see Appendix A):

Theorem 5. GHZ_n *admits a QC-gap if* $n \geq 2$. *(Proof.)*

An interesting (though somewhat unclear) question is how "big" the QC-gap between GHZ_n and QGHZ_n is. In other words, what amount of classical communication is needed to "bridge" the gap? It is known, for instance, that the QC-gap in Bell is "small" in the sense that it only takes a single line of classical communication to simulate all the correlations of QBell [40]. That is, the Bell DAG with communication (23) can reproduce all the correlations of QBell. But how many lines are needed in GHZ_n to simulate all the correlations in QGHZ_n?

Denote by $\text{GHZ}_{n,m}$ the mDAG GHZ_n but supplemented with m "lines" of communication. More formally, $\text{GHZ}_{n,m}$ is the DAG $(V \cup \{\Lambda\}, E)$ where $V = \{\mathtt{X}_1, \ldots, \mathtt{X}_n, \mathtt{Y}_1, \ldots, \mathtt{Y}_n\}$, for all $\mathtt{X}_i \in V$ it holds that $\mathtt{pa}(\mathtt{X}_i) = \varnothing$, and for all $\mathtt{Y}_i \in V$ it holds that $\mathtt{lpa}(\mathtt{Y}_i) = \{\Lambda\}$ and $|\mathtt{vpa}(\mathtt{Y}_i)| = m + 1$. By this definition, $\text{GHZ}_{n,0} = \text{GHZ}_n$. That is, GHZ_n has zero lines of communication.

In terms of QC-gaps, we have the following theorem, which is the content of [41] (though our proof is considerably simpler).[6]

Theorem 6. $\text{GHZ}_{n,m}$ *admits a QC-gap if* $n > m + 1$. *(Proof.)*

We can of course generalize further, and consider the qDAG stemming from $\text{GHZ}_{n,m}$, namely $\text{QGHZ}_{n,m}$. Of course, $\text{GHZ}_{n,m}$ admits a QC-gap relative to $\text{QGHZ}_{n,m}$ for all $m \geq 0$ because $\text{GHZ}_{n,m}$ admits a QC-gap relative to $\text{QGHZ}_{n,0} = \text{QGHZ}_n$.

[5] Though historically Shimony and Mermin had a role to play here too [38, 39].

[6] The author thanks R. Spekkens and M. M. Ansanelli for help with this proof.

3 Computation, Complexity, and Advantage

In this section, we define what it means "to compute", "to be effectively computable", and "to be efficiently computable". We also discuss the circuit based model of computation, the complexity class NC, the distinguished subclass NC^0, which corresponds to the set of decision problems solvable by "shallow circuits", and the one-way quantum computer. We close with two definitions of "quantum advantage" and a discussion on how one might witness the advantage of quantum shallow circuits with QC-gaps.

3.1 Computability and Complexity

The modern definition of "computer" was put forth by Turing [42]. His idea rests on the notion of a *Turing machine*, which Turing envisaged as follows [43]:

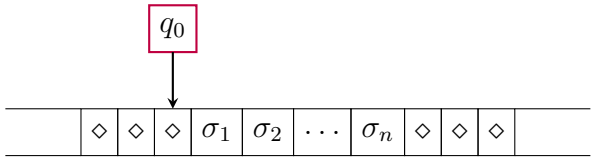
> [A Turing machine has] an unlimited memory capacity obtained in the form of an infinite tape marked out into squares, on each of which a symbol could be printed. At any moment there is one symbol in the machine; it is called the scanned symbol. The machine can alter the scanned symbol, and its behavior is in part determined by that symbol, but the symbols on the tape elsewhere do not affect the behavior of the machine. However, the tape can be moved back and forth through the machine, this being one of the elementary operations of the machine. Any symbol on the tape may therefore eventually have an innings.

Formally, a *deterministic Turing machine* (DTM) $\mathfrak{T}$ is a tuple $(Q, \Sigma, \Gamma, \delta)$, where Q is a finite and nonempty set of *states*, Σ is the *input alphabet*, $\Gamma \supseteq \Sigma \cup \{\diamond\}$ is the *tape alphabet* (with $\diamond$ the *blank symbol*), and δ is the deterministic *transition function*, which bears the form

$$\delta : Q \backslash \{q_{\mathtt{A}}, q_{\mathtt{R}}\} \times \Gamma \to Q \times \Gamma \times \{\triangleleft, \triangleright\}. \tag{34}$$

The state set Q contains three distinguished states: the *initial state* q_0, the *accept state* $q_{\mathtt{A}}$, and the *reject state* $q_{\mathtt{R}}$. Together, $q_{\mathtt{A}}$ and $q_{\mathtt{R}}$ constitute the *halting states* and satisfy $q_{\mathtt{A}} \neq q_{\mathtt{R}}$. Note, since the domain of δ excludes the halting states $q_{\mathtt{A}}$ and $q_{\mathtt{R}}$, $\mathfrak{T}$ *halts* if and only if it transitions to a halting state.

Following Turing, one should picture an initialized DTM $\mathfrak{T}$ as follows:



(35)

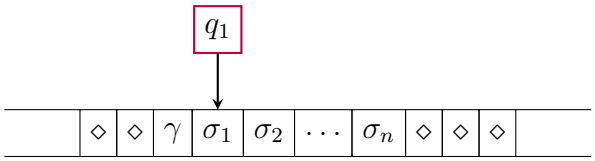
Here, the *finite state control* (purple box) is initialized in the state q_0, and the two-way infinite tape contains the string $x = \sigma_1\sigma_2 \dots \sigma_n \in \Sigma^*$, with one letter $\sigma_i \in \Sigma$ per cell, which encodes the computation to be done. Every other cell of the tape contains the blank symbol $\diamond$. The *tape head* of $\mathfrak{T}$ (black arrow) is initialized so that it points to the blank cell directly to the left of the first symbol in x, namely σ_1. After the initialization, $\mathfrak{T}$ starts taking *steps*, which are calls to the transition function δ. For example, the first step could be $\delta(q_0, \diamond) = (q_1, \gamma, \triangleright)$, which, in the canonical interpretation, means:

(i) the state of $\mathfrak{T}$ transitions from q_0 to some $q_1 \in Q$,

(ii) the tape head overwrites $\diamond$ with a (not necessarily new) symbol $\gamma \in \Gamma$,

(iii) the tape head moves one cell to the right (as defined by the symbol $\triangleright$).

Pictorially, this new configuration is:



(36)

We have laboured through this definition, and in particular its canonical physical interpretation, to emphasize the following point. It is useful, though not strictly necessary, to endorse Turing's physical interpretation in order to justify why a DTM is what we formally mean by a "computer". Indeed, the physical interpretation makes it plain that the mathematical structure of a DTM *at least* depicts a human working by rote, scribbling symbol after symbol with pencil and paper. Thus, whatever we can compute with pencil and paper, a DTM can compute too.[7] Of course, we need not be so anthropomorphic, but that the DTM *at least* captures the image of a human working

[7]Incidentally, this means that DTM's are a mathematical structure that naturally encode the process of "doing math". It is for this self-referential reason that Gödel's incompleteness theorems are easily castable in a DTM guise [44].

by rote means a DTM *at least* has something to do with most garden-variety notions of "computation". In fact, it is accepted today that to be *computable*, or perhaps more precisely, to be *effectively computable* is to be solvable on a DTM [45]. Of course, what exactly we mean by "solvable" is still unclear.

The above description makes it evident that every input $x \in \Sigma^*$ causes a DTM $\mathfrak{T}$ to accept, reject, or neither. One can therefore consider the set $L(\mathfrak{T}) = \{x \in \Sigma^* \mid \mathfrak{T}(x) \text{ accepts}\}$, which is the *language computed by* $\mathfrak{T}$.

Historically, complexity theory has dealt with *language decision problems*, which ask if a string $x \in \Sigma^*$ is in a set $L \subseteq \Sigma^*$ or not (L is called a *language*). For example, "is 42 composite?" is a simple decision problem. If $L \subseteq \Sigma^*$ is a language, then L is *recognizable* (also known as *recursively enumerable*) if and only if there exists a DTM $\mathfrak{T}$ such that $L(\mathfrak{T}) = L$. This is to be contrasted with the notion of *decidable*: L is *decidable* if and only if there exists a DTM $\mathfrak{T}$ such that $L(\mathfrak{T}) = L$ and $\mathfrak{T}(x)$ rejects for all $x \in \Sigma^* \backslash L$. That recognizable but undecidable languages exist is nontrivial and profound [42, 45].

Of course, Turing machines can compute more than languages. They can also compute certain functions $f : \Sigma^* \to \Sigma^*$. Specifically, DTM's define the set of functions that are *effectively computable*: a function $f : \Sigma^* \to \Sigma^*$ is *effectively computable* if and only there exists a DTM $\mathfrak{T}$ such that for all $x \in \Sigma^*$ the input $\langle x \rangle$ implies $\mathfrak{T}$ halts after a finite number of steps with $f(x)$ written on its tape. In this sense, Turing machines *define* computability. Notice, this notion subsumes language decision problems, since determining if $x \in L$ or $x \notin L$ is equivalent to computing the Boolean function $f_L : \Sigma^* \to \{0, 1\}$, where $f_L(x) = 1$ if $x \in L$ and $f_L(x) = 0$ otherwise.

Given that DTMs can evaluate effectively computable functions, we can now ask a Turing machine to solve more complicated computational tasks. Two important examples are *search* and *sampling problems*.

A *search problem* R is a collection of nonempty sets $(A_x)_{x \in \Sigma^*}$, where each $A_x \subseteq \Sigma^{p(|x|)}$ with p a fixed polynomial function and $|x|$ denotes the length of the string x.[8] On the other hand, a *sampling problem* S is a collection of probability distributions $(\mathscr{D}_x)_{x \in \Sigma^*}$, where each $\mathscr{D}_x$ is a distribution over $\Sigma^{p(|x|)}$ with p again a fixed polynomial function. How a DTM "solves" these is not obvious (especially the distribution problem). The answer, however, ultimately rests in the robustness of the Turing machine model.

[8] We require p to be a polynomial so that the Turing machine attempting to solve R has the potential to halt in a polynomial amount of time. As we will talk about shortly, halting in a polynomial amount of time is more or less the definition of "efficient".

Suppose, for instance, that we modify the transition function (34) so that

$$\delta : Q\backslash\{q_{\mathtt{A}}, q_{\mathtt{R}}\} \times \Gamma \to 2^{Q\times\Gamma\times\{\triangleleft,\triangleright\}}. \tag{37}$$

Then at each step this *nondeterministic* Turing machine (NTM) will map to a set of possible transitions (hence "nondeterministic"). In particular, there could be a degree of internal randomness that chooses which transition to make, in which case the NTM is a *probabilistic* Turing machine (PTM). It then makes sense to talk about solving problems probabilistically.

For instance, a PTM $\mathfrak{T}$ is said to *solve a search problem* $R = (A_x)_{x\in\Sigma^*}$ if and only if for all $\epsilon > 0$ and for all $x \in \Sigma^*$ the input $\langle x, 0^{1/\epsilon}\rangle$ implies[9]

$$\Pr_{\mathtt{X}\sim\mathscr{D}(\mathfrak{T})}(\mathtt{X} \in A_x) \geq 1 - \epsilon, \tag{38}$$

where the probability distribution $\mathscr{D}(\mathfrak{T})$ and the associated random variable $\mathtt{X}$ arise from the internal randomness of the PTM $\mathfrak{T}$. Importantly, to solve R to any small $\epsilon > 0$, it suffices to contrive a PTM that solves R to any $\tilde{\epsilon}$ for which $\epsilon < \tilde{\epsilon} < 1/2$ (complexity theorists usually choose $\tilde{\epsilon} = 1/3$). This is trivial, because as long as the PTM is more likely than not to get the right answer, then rerunning the PTM $O(1/\epsilon)$ times and taking the "majority vote" will yield the right answer to within any desired accuracy [45].

On the other hand, $\mathfrak{T}$ is said to *solve a sampling problem* $S = (\mathscr{D}_x)_{x\in\Sigma^*}$ if and only if for all $\epsilon > 0$ and for all $x \in \Sigma^*$ the input $\langle x, 0^{1/\epsilon}\rangle$ implies

$$\|\mathscr{D}(\mathfrak{T}) - \mathscr{D}_x\| \leq \epsilon, \tag{39}$$

where $\|\cdot\|$ is the total variation distance [17, 45]. Unlike in the case of search problems, to solve a sampling problem within any $\epsilon > 0$ it does *not* suffice to solve it within any $\tilde{\epsilon}$ such that $\epsilon < \tilde{\epsilon} < 1/2$. This follows because of the nature of the problem: repeated trials of the PTM $\mathfrak{T}$ *define* the distribution $\mathscr{D}(\mathfrak{T})$, so repeated trials cannot possibly reduce the distance between $\mathscr{D}(\mathfrak{T})$ and a target distribution $\mathscr{D}_x \in S$. In brief: for sampling, ϵ matters.

Interestingly, a DTM can solve search and sampling problems as well, because every NTM can be simulated by a DTM (though the simulation will generally take an exponential number of steps). This is an instance of the more general *Church-Turing thesis* [46]. Consequently, where the advantage

[9]The additional $0^{1/\epsilon}$ input is a unary way to tell the Turing machine "be within ϵ of the right answer". Of course, the computation takes longer as $\epsilon \to 0$ because the Turing machine must take more steps to read and hence reach the desired accuracy.

of nondeterminism, and more specifically probability, comes in (if there is one at all) is in reducing the number of steps it takes to get an answer. This brings us from the field of computability theory, concerned with what is effectively computable, to the field of complexity theory, concerned with what is *efficiently computable*.

According to *Cobham's thesis*, a function is *efficiently computable* if and only if it can be computed on a Turing machine in a time polynomial in the size of the input [46, 47]. This is intimately tied to the *extended Church-Turing thesis*, which basically posits that any "realistic" model of computation is efficiently simulable on a probabilistic Turing machine [46].

Unlike when reasoning about computability theory, when it comes to efficiently the differences in DTMs, PTMs, and NTMs really seem to make a difference. For example, the set of efficiency decidable languages on a DTM equals P, the set of efficiently decidable languages (with acceptance probability at least $2/3$) on a PTM equals BPP, the set of efficiently decidable languages on an NTM equals NP, but the exact relationship between these three classes is unknown. Indeed, no relationship between BPP and NP is currently known [44], and it is literally a million dollar question if at least one of the following obvious containments is strict or not: $\mathsf{P} \subseteq \mathsf{BPP}$ and $\mathsf{P} \subseteq \mathsf{NP}$ [48].[10]

3.2 Boolean Circuits

Turing's model of computation is *uniform*, meaning a Turing machine's function is irrespective of the size of the input. Circuits, on the other hand, are *non-uniform*, meaning the circuit is a function of the size of the input. In other words, the circuit that solves an n bit instance of a problem will not in general solve its $n+1$ bit generalization.[11] This distinction rests on the idea of a *Boolean function*, which is simply an n-ary map $f : \{0,1\}^n \to \{0,1\}$.

Canonical (and trivial) examples of Boolean functions include the unary ($n = 1$) NOT operation $\neg : \{0,1\} \to \{0,1\}$ as well as the binary ($n = 2$) AND and OR operations $\wedge_2, \vee_2 : \{0,1\}^2 \to \{0,1\}$. Of course, these binary operations can be extended to accommodate a larger arity:

$$\wedge_n : \{0,1\}^n \to \{0,1\}, \qquad \wedge_n(x_1, \ldots, x_n) \coloneqq x_1 \wedge \cdots \wedge x_n, \tag{40}$$

$$\vee_n : \{0,1\}^n \to \{0,1\}, \qquad \vee_n(x_1, \ldots, x_n) \coloneqq x_1 \vee \cdots \vee x_n. \tag{41}$$

[10] Indeed, $\mathsf{P} \neq \mathsf{BPP}$ implies $\mathsf{P} \neq \mathsf{NP}$ [44].

[11] Without loss of generality, we hereafter use the binary alphabet $\Sigma = \{0,1\}$.

However, $\wedge_n$ on $n+1$ bits is ill-defined, and that is analogically why the circuit model of computation is non-uniform. We mend this issue with the notion of a *family of Boolean functions*, which is a sequence $f = (f_n)_{n\in\mathbb{N}}$, where each $f_n : \{0,1\}^n \to \{0,1\}$ is an n-ary Boolean function. For instance, defining $\wedge := (\wedge_n)_{n\in\mathbb{N}}$, it becomes clear that $\wedge(x_1,\ldots,x_n) = \wedge_n(x_1,\ldots,x_n)$ because that is the only function in the family $\wedge$ with arity n. Thus for the family f, we have $f(x) := f_{|x|}(x)$, where $|x|$ denotes the length of the input.

Any finite collection of Boolean functions and families of Boolean functions constitutes a *basis*. A basis, therefore, is a finite object, but the elements therein need not be finite (in particular, families of Boolean functions are infinitely large). There are therefore two distinguished types of bases: those with *bounded fan-in* (i.e., those which contain no family of Boolean functions) and those with *unbounded fan-in* (i.e., those which contain at least one family of Boolean functions). In this essay we only care about bounded fan-in bases. Therefore, by "basis" we hereafter mean "bounded fan-in basis".

A particularly distinguished basis is the *standard basis* $B_0 := \{\wedge_2, \vee_2, \neg\}$. This basis is universal in the sense that any n-ary Boolean function $f : \{0,1\}^n \to \{0,1\}$ can be expressed using only elements in B_0. This follows because every Boolean function admits a disjunctive normal form (DNF) [45]. For the same reason, any n-ary *vector-valued* Boolean function $f : \{0,1\}^n \to \{0,1\}^m$, $m \geq 1$, can be expressed using only elements in B_0. The DNF of a vector-valued Boolean function can be construed as an example of a Boolean circuit over the basis B_0.

Generally speaking, a Boolean circuit over a basis B is a graphical representation of a sequence of Boolean functions in B. More precisely, if B is a basis, then a *Boolean circuit* over B with n inputs and m outputs is a DAG $\mathcal{C}_n = (V, E)$ in which a node of in-degree zero is either one of n inputs or a 0-ary vector-valued Boolean function (i.e., a *Boolean constant*) in B, a node of in-degree $k \geq 1$ is a k-ary vector-valued Boolean function in B, and a node of out-degree zero is one of m outputs. A Boolean circuit is therefore an elaborate composition of primitive Boolean functions in the basis B, and hence the circuit itself corresponds to a vector-valued Boolean function $\mathcal{C}_n : \{0,1\}^n \to \{0,1\}^m$.

That the DNF of a vector-valued Boolean function is a Boolean circuit proves that every vector-valued Boolean function can be represented as a Boolean circuit. In this sense, if $f = (f_n)_{n\in\mathbb{N}}$ is a family of vector-valued Boolean functions, then there is at least one *family of Boolean circuits* over B_0 that *computes* f. More particularly, the circuit family $(\mathcal{C}_n)_{n\in\mathbb{N}}$ *computes*

f if and only if for all $x \in \{0,1\}^*$ the circuit $\mathcal{C}_{|x|}$ computes the restriction of f to strings of length $|x|$. In other words, for every $x \in \{0,1\}^*$ it holds that $\mathcal{C}_{|x|}(x) = f(x) = f_{|x|}(x)$.

Of course, the circuit family $(\mathcal{C}_n)_{n\in\mathbb{N}}$ computing the function family $(f_n)_{n\in\mathbb{N}}$ need not scale in any sort of "uniform" way, meaning the architecture of $\mathcal{C}_n$ may look totally different from the architecture of $\mathcal{C}_{n+1}$.

Of the many important architectural features of a Boolean circuit $\mathcal{C}_n = (V, E)$ over the basis B, two are distinguished in terms of measuring complexity. The first is the *size* of $\mathcal{C}_n$, which is the number of non-input and non-output gates in V—that is, $|\{v \in V \mid v \in B\}|$. The second is the *depth* of $\mathcal{C}_n$, which is the length of the longest directed path in the graph (V, E).

Roughly speaking, the depth of a circuit is a measure of how parallelizable the function it represents is. It is easy to prove that the size of $\mathcal{C}_n$ is at most exponential in the depth, since the worst case is that the circuit looks like a k-ary tree, where k is the maximum fan-in of the gates in the basis B [49].

What size is reasonable? There are 2^n Boolean functions $f : \{0,1\}^n \to \{0,1\}$, so there are 2^{nm} vector-valued Boolean functions $f : \{0,1\}^n \to \{0,1\}^m$. Since we are interested in functions that are efficiently computable, we require m to be polynomial in n, otherwise it would take a DTM more than polynomial time to write the answer. Consequently, the functions we care about take the form $f : \{0,1\}^n \to \{0,1\}^{p(n)}$, where p is a polynomial. The DNF of a function like this requires an exponential number of gates in B_0 [45]. Therefore, exponential size Boolean circuits can compute all the functions we might care about. That is not interesting, so it makes sense to restrict to *polynomial size* Boolean circuits.

The set of languages decidable by polynomial size Boolean circuits is $\mathsf{P/poly}$. Here "poly" means "polynomial advice". Canonically, *advice* is defined as an additional input to a Turing machine whose length is a function of the length of the input. *Polynomial advice*, therefore, is advice whose length is polynomial in the length of the input. That the languages efficiently decidable by a DTM with polynomial advice ($\mathsf{P/poly}$) equals the languages decidable by polynomial size Boolean circuits is an elementary theorem [44, 45].

Incidentally, advice is extremely powerful. For one thing, *Adelman's theorem* proves $\mathsf{BPP} \subseteq \mathsf{P/poly}$, which means that non-uniformity is at least as powerful as randomness [44, 50]. But one can go farther: Turing machines with advice, and hence polynomial size Boolean circuits, can solve undecidable problems like the unary halting problem [45, 47]. Thus, Adelman's

result is strict: $\mathsf{BPP} \subsetneq \mathsf{P/poly}$. Polynomial size Boolean circuits, therefore, constitute an immensely powerful computational model that transcends the canonical Turing machine.

3.3 Classical Shallow Circuits and NC^0

To render the Boolean circuit model more reasonable, we must constrain it beyond just polynomial size circuits. One particularly interesting constraint is to polynomial size circuits over B_0 with depth at most $O(\log^i n)$ for some $i \geq 0$. Furthermore, we require that families of such circuits $(\mathcal{C}_n)_{n\in\mathbb{N}}$ be *uniformly generated*. This means that for each family there exists a DTM $\mathfrak{T}$ such that the unary input $\langle 0^n \rangle$ implies $\mathfrak{T}$ outputs a description of $\mathcal{C}_n$ using at most $O(\log n)$ tape cells (also known as *log-space*) [45]. Incidentally, this log-space uniformity constraint guarantees that the generated circuits are at most polynomial in size.

Altogether, the set of languages decidable by such circuits is NC^i, where "NC" stands for "Nick's class" after Nick Pippenger [51, 52]. Formally, NC is the set of languages decidable by uniformly generated polynomial size and polylogarithmic depth circuits:

$$\mathsf{NC} \coloneqq \bigcup_{i\geq 0} \mathsf{NC}^i. \tag{42}$$

Intuitively, NC corresponds to the set of languages that are efficiently decidable by a parallel computer [45]. This means that a language $L \in \mathsf{NC}$ if and only if deciding L can be "broken up" into smaller subproblems that need not "talk" to each other during much of the calculation. It follows, therefore, that $\mathsf{NC} \subseteq \mathsf{P}$, because a polylogarithmic number of parallel computations can be simulated one after the other in polynomial time [45]. However, it is unknown if $\mathsf{NC} \subseteq \mathsf{P}$ is strict. That is, it is unknown if every efficiently decidable language is also parallelizable [49, 52].

Importantly, whereas all polynomial size and polylogarithmic depth circuits are canonically defined over the standard basis B_0, they are in fact robust to changes in this basis. More precisely, if $B \cup B_0$ is a basis and $(\mathcal{C}_n)_{n\in\mathbb{N}}$ is a family of polynomial size and depth $O(\log^i n)$ circuits over $B \cup B_0$, then there exists a family of polynomial size and depth $O(\log^i n)$ circuits $(\widetilde{\mathcal{C}}_n)_{n\in\mathbb{N}}$ over B_0 such that for all $x \in \{0,1\}^*$ it holds that $\mathcal{C}_{|x|}(x) = \widetilde{\mathcal{C}}_{|x|}(x)$ [49]. Therefore, when reasoning about polynomial size and polylogarithmic depth

circuits, the basis we choose is incidental. In particular, the maximum fan-in of the basis is immaterial insofar as it is finite.

Of course, circuits of depth $O(\log^{i+1} n)$ are feasibly more powerful than circuits of depth $O(\log^{i} n)$. That is, $\mathsf{NC}^1 \subseteq \mathsf{NC}^2 \subseteq \cdots$, though it is outstanding if any of these containments is strict [45, 49]. That said, it is relatively straightforward to prove $\mathsf{NC}^0 \subsetneq \mathsf{NC}^1$ because *constant depth* or *shallow* circuits cannot decide if the majority of their input was 0 or 1 [49]. Thus, shallow circuits constitute a very restricted computational model.

Nevertheless, shallow circuits are interesting for a variety of reasons. Generally speaking, unconditional impossibility results are difficult to come by in theoretical computer science. However, shallow circuits afford many such results, and therefore have been involved in several of the early successes of complexity theory [18, 53]. Shallow circuits are also of great practical interest. This is because shallow circuits involve little "communication" between the internal gates, and this makes them less prone to compounding errors. Incidentally, this is also true of *quantum shallow circuits* (Sec. 3.4) [18].

As shallow circuits are very restricted, it is natural to wonder what they can compute. In particular, what search and sampling problems can they solve? This question is prima facie unclear because in Sec. 3.1 "solving" search and sampling problems involved a degree of randomness. With shallow circuits, however, there is no inherent source of randomness to exploit. Fortunately, this issue is easily mended: allow the shallow circuit a source of *polynomial randomized advice*, which is denoted rpoly. In other words, to each of the polynomial number of gates in a shallow circuit $\mathcal{C}_n$, provide a constant number of bits from a polynomial size bit string that is drawn from a distribution $\mathscr{R}$. Altogether, then, the shallow circuit is supplemented with a polynomial length advice string $\Lambda \sim \mathscr{R}$ that the circuit can use to compute.[12]

With randomized advice, it makes sense to talk about probabilistic outcomes. For sake of generality, we can imagine that each input bit $\sigma_i \in \{0, 1\}$ comes from conditioning on the valuation $\mathtt{X}_i = \sigma_i$ of some $\mathscr{D}_i$-random variable $\mathtt{X}_i$. Then, $\mathtt{X} = (\mathtt{X}_1, \ldots, \mathtt{X}_n)$ is a joint random variable over $\{0, 1\}^n$, and the valuation $\mathtt{X} = x \in \{0, 1\}^n$ corresponds to an n bit circuit input. Furthermore, due to the randomized advice Λ, the output string is also some joint random variable $\mathtt{Y} = (\mathtt{Y}_1, \ldots, \mathtt{Y}_{p(n)})$ over $\{0, 1\}^{p(n)}$ for some polynomial

[12] Alternatively, one can think of randomized advice as an additional input to the circuit, rather than as providing a constant number of bits to each gate in the circuit.

p. Therefore, $\mathtt{X}$ and $\mathtt{Y}$ constitute a joint $\mathscr{D}$-random variable, where the PMF is the marginalized distribution

$$\Pr_{\mathtt{X},\mathtt{Y}\sim\mathscr{D}}(\mathtt{X},\mathtt{Y}) = \sum_{\Lambda=\lambda} f(\mathtt{X},\mathtt{Y},\lambda)\Pr_{\Lambda}(\lambda) \tag{43}$$

for some function f.

Physically speaking, the wires in a classical circuit dictate the cause and effect relationships between the various gates. Consequently, the PMF (43) must factorize according to the directed local Markov property:

$$\Pr_{\mathtt{X},\mathtt{Y}\sim\mathscr{D}}(\mathtt{X},\mathtt{Y}) = \sum_{\Lambda=\lambda}\prod_{i=1}^{n}\prod_{j=1}^{p(n)}\Pr_{\mathtt{Y}_j|\mathtt{pa}(\mathtt{Y}_j)}(\mathtt{Y}_j \mid \mathtt{pa}(\mathtt{Y}_j))\Pr_{\mathtt{X}_i}(\mathtt{X}_i)\Pr_{\Lambda}(\lambda). \tag{44}$$

Here we assume that the input random variables $\mathtt{X}_1,\ldots,\mathtt{X}_n$ have no parents and do not influence the advice. The Markovian parents of each output random variable $\mathtt{Y}_j$ is constrained by the following lemma [11, 54].

Lemma 7. *If $\mathcal{C}_{n+r} : \{0,1\}^{n+r} \to \{0,1\}^{p(n)}$ is a classical shallow circuit with $r = \mathrm{poly}(n)$ bits of randomized advice Λ, depth d, and fan-in k, then $|\mathtt{pa}(\mathtt{Y}_j)\backslash\{\Lambda\}| \leq k^d$ for all $\mathtt{Y}_j \in \{\mathtt{Y}_1,\ldots,\mathtt{Y}_{p(n)}\}$.* *(Proof.)*

In other words, each output bit of a classical shallow circuit can depend on at most k^d non-advice input bits. Lemma 7 immediately entails that classical shallow circuits output distributions that are compatible with $\mathrm{GHZ}_{p(n),k^d}$:

Theorem 8. *Let $\mathcal{C}_{n+r} : \{0,1\}^{n+r} \to \{0,1\}^{p(n)}$ be an $n+r$ bit classical shallow circuit with $r = \mathrm{poly}(n)$ bits of randomized advice, depth d, and fan-in k. Then $\mathscr{D}(\mathcal{C}_{n+r})$ is compatible with $\mathrm{GHZ}_{p(n),k^d}$.*

Consequently, every classical shallow circuit can be thought of as a particular GHZ experiment with a fixed amount of nearest neighbor communication. This theorem is central to our results in Sec. 3.5.

As we vary the valuations $\mathtt{X} = x$, the *output distribution of the circuit* given the input $\langle x\rangle$ is the conditional distribution

$$\Pr_{\mathtt{Y}|\mathtt{X}\sim\mathscr{D}(\mathcal{C}_{n+r})}(\mathtt{Y} \mid \mathtt{X}) \coloneqq \frac{\Pr_{\mathtt{X},\mathtt{Y}\sim\mathscr{D}}(\mathtt{X},\mathtt{Y})}{\sum_{\mathtt{X}_1=\sigma_1}\cdots\sum_{\mathtt{X}_n=\sigma_n}\Pr_{\mathtt{X},\mathtt{Y}\sim\mathscr{D}}(\mathtt{Y}_1,\ldots,\mathtt{Y}_{p(n)},\sigma_1,\ldots,\sigma_n)}, \tag{45}$$

Formally, it now makes sense to talk about shallow circuits solving problems probabilistically. But what qualifies as "solving"? The following definitions are our circuit adaptations of the corresponding ones in Sec. 3.1, which were for a BPP Turing machine.

Definition 5 ($\mathsf{FNC}^0/\mathsf{rpoly}_\epsilon$)**.** Fix $\epsilon \in (0, \frac{1}{2})$. The complexity class $\mathsf{FNC}^0/\mathsf{rpoly}_\epsilon$ consists of all search problems $R = (A_x)_{x\in\{0,1\}^*}$ for which there exists a uniformly generated family of classical shallow circuits $(\mathcal{C}_{n+r})_{n\in\mathbb{N}}$ with $r = \mathrm{poly}(n)$ bits of randomized advice such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies

$$\Pr_{\mathtt{Y}|\mathtt{X}\sim\mathscr{D}(\mathcal{C}_{|x|+r})}(\mathtt{Y} \in A_x \mid \mathtt{X} = x) \geq 1 - \epsilon. \tag{46}$$

Unlike before, to solve $R \in \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon$ to any small $\epsilon > 0$ it does not suffice to find a shallow circuit family that solves R to any $\tilde{\epsilon}$ for which $\epsilon < \tilde{\epsilon} < 1/2$. This follows from Theorem 2 in [11]:

Theorem 9. *There exist* $\epsilon_1 < \epsilon_2 < \frac{1}{2}$ *such that* $\mathsf{FNC}^0/\mathsf{rpoly}_{\epsilon_1} \subsetneq \mathsf{FNC}^0/\mathsf{rpoly}_{\epsilon_2}$.

Intuitively, a result like this is expected because shallow circuits cannot process a "majority vote" [49]. Thus, repeated trials on a shallow circuit cannot amplify an $\tilde{\epsilon} > 0$ to some positive $\epsilon < \tilde{\epsilon}$ like in the Turing machine model of computation. The story is probably different for shallow circuits with unbounded fan-in, however [53].

Importantly, Theorem 9 implies randomized advice empowers FNC^0:

Corollary 10. *There exists* $\epsilon \in (0, \frac{1}{2})$ *for which* $\mathsf{FNC}^0 \subsetneq \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon$.

The proof is trivial.

Incidentally, the same is not true for the set of language decision problems decidable by shallow circuits with randomized advice:

Lemma 11. *For all* $\epsilon \in (0, \frac{1}{2})$ *it holds that* $\mathsf{NC}^0 = \mathsf{NC}^0/\mathsf{rpoly}_\epsilon$. *(Proof.)*

Therefore, randomized advice does not empower shallow circuits for language decision problems.

Consider now the complexity class

$$\mathsf{FNC}^0/\mathsf{rpoly} \coloneqq \bigcap_{\epsilon\in(0,\frac{1}{2})} \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon. \tag{47}$$

In words, $\mathsf{FNC}^0/\mathsf{rpoly}$ consists of all search problems $R = (A_x)_{x\in\{0,1\}^*}$ for which for all $\epsilon > 0$ there exists a family of classical shallow circuits $(\mathcal{C}_{n+r_\epsilon})_{n\in\mathbb{N}}$ with $r_\epsilon = \mathrm{poly}(n, 1/\epsilon)$ bits of randomized advice, depth d_ϵ, and fan-in k_ϵ such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies Eq. (46).[13] Thus, $\mathsf{FNC}^0/\mathsf{rpoly}$ consists of the "easiest" search problems solvable by shallow circuits because they can be solved to any precision. We do not know if $\mathsf{FNC}^0 = \mathsf{FNC}^0/\mathsf{rpoly}$, but conjecture that this is false:

Conjecture 12. $\mathsf{FNC}^0 \subsetneq \mathsf{FNC}^0/\mathsf{rpoly}$.

Sampling problems are defined similarly to search problems.

Definition 6 ($\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon$)**.** The complexity class $\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon$ consists of all sampling problems $S = (\mathscr{D}_x)_{x\in\{0,1\}^*}$ for which there exists a uniformly generated family of classical shallow circuits $(\mathcal{C}_{n+r})_{n\in\mathbb{N}}$ with $r = \mathrm{poly}(n)$ bits of randomized advice such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies

$$\left\|\mathscr{D}(\mathcal{C}_{|x|+r}) - \mathscr{D}_x\right\| \leq \epsilon. \tag{48}$$

The analogue of Lemma 9 for search problems is the following lemma, which actually follows from Lemma 9:

Lemma 13. *There exist* ϵ_1, ϵ_2 *such that* $\mathsf{SampNC}^0/\mathsf{rpoly}_{\epsilon_1} \subsetneq \mathsf{SampNC}^0/\mathsf{rpoly}_{\epsilon_2}$. *(Proof.)*

Therefore, random advice empowers shallow circuits for sampling problems (though this is hardly surprising):

Corollary 14. *There exists* $\epsilon > 0$ *such that* $\mathsf{SampNC}^0 \subsetneq \mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon$.

Similar to search problems, we define the "easiest" sampling problems solvable on a shallow circuit by

$$\mathsf{SampNC}^0/\mathsf{rpoly} \coloneqq \bigcap_{\epsilon>0} \mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon. \tag{49}$$

Formally, $\mathsf{SampNC}^0/\mathsf{rpoly}$ is the set of all sampling problems $S = (\mathscr{D}_x)_{x\in\{0,1\}^*}$ for which for all $\epsilon > 0$ there exists a family of classical shallow circuits $(\mathcal{C}_{n+r_\epsilon})_{n\in\mathbb{N}}$ with $r_\epsilon = \mathrm{poly}(n, 1/\epsilon)$ bits of randomized advice, depth d_ϵ, and fan-in k_ϵ such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies Eq. (48).

When it comes to empowerment, random advice empowers shallow circuits even for the easiest problems (again, this is not too surprising):

[13] Here, subscripts denote a potential dependence on ϵ.

Lemma 15. $\mathsf{SampNC}^0 \subsetneq \mathsf{SampNC}^0/\mathsf{rpoly}$. *(Proof.)*

Here, the separating problem is to essentially mimic the roll of a large dice.

3.4 Quantum Shallow Circuits and QNC^0

An n-partite quantum state ρ_n can be measured by n parties with the hope of sampling from some distribution $\mathscr{D}_n$ over n bits. In this way, measurements on a family of quantum states $(\rho_n)_{n\in\mathbb{N}}$ can feasibly yield statistics that solve a sampling problem $(\mathscr{D}_n)_{n\in\mathbb{N}} \in \mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon$ for some $\epsilon > 0$. But what quantum states are allowed? To make the comparison to classical shallow circuits fair, it is natural to require ρ_n to be realizable on a *quantum shallow circuit*, which is a restricted type of *quantum computer*.

Although most have an intuitive understanding of what is meant by "quantum computer", a formal definition of this term is actually quite elusive. In fact, in Nielsen and Chuang's textbook [55], which is canonically regarded as *the* authoritative reference for introductory quantum computing, "quantum computer" is never defined. As underscored in [56], the literature tends to dance around several different types of definitions of "quantum computer", from quantum Turing machines [57] to adiabatic quantum processors [58].

The circuit based model, however, is perhaps the most widespread, which Yao proved to be equivalent to the quantum Turing machine [59]. Here the image of a quantum computer is as an array of initially separated qubits that pass through a network of basic quantum operations [55, 60]. The set of such basic quantum operations is called a *quantum gate set* and is akin to the basis of a classical Boolean circuit.

Like we did for classical Boolean circuits, we restrict to quantum gate sets with bounded fan-in. Incidentally, the no-cloning theorem implies that quantum gate sets must also have bounded fan-out.

In this essay we use the "Clifford + T" gate set, which includes the single-qubit gates

$$H = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad S = \begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}, \quad \text{and} \quad T = \begin{pmatrix} 1 & 0 \\ 0 & e^{i\pi/4} \end{pmatrix} \tag{50}$$

as well as the two-qubit controlled-Z gate $\mathrm{CZ} = |0\rangle\langle 0| \otimes \mathbb{1}_2 + |1\rangle\langle 1| \otimes Z$, where $\mathbb{1}_n$ is the $n \times n$ identity matrix and

$$Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{51}$$

is the Pauli Z gate. We also require the Pauli X and $R_z(\theta)$ rotation gates:

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad R_z(\theta) = \begin{pmatrix} e^{-i\theta/2} & 0 \\ 0 & e^{i\theta/2} \end{pmatrix}. \tag{52}$$

Given the Clifford+T gate set, we define a *layer* U as a finite tensor product of operators in Clifford+T. Thus a layer is a unitary operator. The *size of the layer* U is the number of Clifford+T gates that compose it. A *quantum circuit* $\mathcal{Q}$ *of depth* d consists of a sequence of d layers $U_1, \ldots, U_d$ multiplied together in reverse order: $\mathcal{Q} = U_d \cdots U_1$. The *size of the quantum circuit* $\mathcal{Q}$ is the sum of the sizes of the layers $U_1, \ldots, U_d$. Altogether, then, a quantum circuit of depth d corresponds to the unitary operator

$$\mathcal{Q} = U_d \cdots U_1 = \prod_{i=1}^{d} \bigotimes_{j(i)} O_j, \tag{53}$$

where O_j is some Clifford+T gate. That every unitary operator can be approximated arbitrarily by a factorization of this sort is the statement that the Clifford+T gate set is *universal*. This is nontrivial and profound [55].

Whereas in a classical circuit the input is a bit string $x = \sigma_1 \ldots \sigma_n \in \{0,1\}^n$, in a quantum circuit the input is a qudit $|x\rangle = |\sigma_1\rangle \otimes \cdots \otimes |\sigma_n\rangle$. Thus, on input $\langle x\rangle$, the output of a quantum circuit $\mathcal{Q}$ is the state $\mathcal{Q}|x\rangle$. To obtain a classical output, we measure this state in the computational basis. Thus, on input $|x\rangle$, the *distribution of a quantum circuit* $\mathcal{Q}$ is

$$\Pr_{\mathtt{Y} \sim \mathscr{D}(\mathcal{Q})}(\mathtt{Y}) = |\langle \mathtt{Y}|\mathcal{Q}|x\rangle|^2 \tag{54}$$

for every valuation $\mathtt{Y} = y \in \{0,1\}^n$. Incidentally, Eq. (54) is identical to

$$\Pr_{\mathtt{Y} \sim \mathscr{D}(\mathcal{Q})}(\mathtt{Y}) = \mathrm{Tr}\big(\rho^{\mathcal{Q}}(x) M_{\mathtt{Y}}\big), \tag{55}$$

where $\rho^{\mathcal{Q}}(x) = \mathcal{Q}|x\rangle\langle x|\mathcal{Q}^\dagger$ is the density operator of the state $\mathcal{Q}|x\rangle$ and $M_{\mathtt{Y}} = |\mathtt{Y}\rangle\langle \mathtt{Y}|$ projects into the $\mathtt{Y}$ subspace. Like for classical circuits, we allow for the possibility that the input to the quantum circuit is decided by a classical random variable $\mathtt{X}$. In this case, the output distribution Eq. (55) becomes the conditional distribution

$$\Pr_{\mathtt{Y}|\mathtt{X} \sim \mathscr{D}(\mathcal{Q})}(\mathtt{Y} \mid \mathtt{X}) = \mathrm{Tr}\big(\rho^{\mathcal{Q}}(\mathtt{X}) M_{\mathtt{Y}}\big). \tag{56}$$

It is the purpose of the remainder of this section to find a set of conditions under which this distribution is compatible with QGHZ_n. Among other conditions, it will turn out that we require $\mathcal{Q}$ to be *shallow*.

Let $\mathcal{Q}_n$ be a quantum circuit on n qubits. Then a *family of quantum shallow circuits* $(\mathcal{Q}_n)_{n\in\mathbb{N}}$ is a family of quantum circuits whose depth d does not scale with n. Similar to classical shallow circuits, we say the family $(\mathcal{Q}_n)_{n\in\mathbb{N}}$ is *uniformly generated* if and only if there exists a DTM $\mathfrak{T}$ such that $\mathfrak{T}(0^n)$ yields a description of $\mathcal{Q}_n$ in log-space. In this way, classical and quantum shallow circuit families are placed on the same footing, and so a comparison between the two becomes fair.

Given the output distributions of a quantum shallow circuit, we can now define the set of sampling and search problems that they can compute. Of course, these are analogous to the classical definitions.

Definition 7 ($\mathsf{SampQNC}^0_\epsilon$)**.** Fix $\epsilon > 0$. The complexity class $\mathsf{SampQNC}^0_\epsilon$ consists of all sampling problems $S = (\mathscr{D}_x)_{x\in\{0,1\}^*}$ for which there exists a uniformly generated family of quantum shallow circuits $(\mathcal{Q}_n)_{n\in\mathbb{N}}$ such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies

$$\left\| \mathscr{D}(\mathcal{Q}_{|x|}) - \mathscr{D}_x \right\| \leq \epsilon. \tag{57}$$

This definition is a refinement of the definition of $\mathsf{SampQNC}^0$ given in [61]. The same is also true for their definition of FQNC^0, which we define as follows.

Definition 8 (FQNC^0_ϵ)**.** Fix $\epsilon \in (0, \frac{1}{2})$. The complexity class FQNC^0_ϵ consists of all search problems $R = (A_x)_{x\in\{0,1\}^*}$ for which there exists a uniformly generated family of quantum shallow circuits $(\mathcal{Q}_n)_{n\in\mathbb{N}}$ such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies

$$\Pr_{\mathtt{Y}|\mathtt{X}\sim\mathscr{D}(\mathcal{Q}_{|x|})}(\mathtt{Y} \in A_x \mid \mathtt{X} = x) \geq 1 - \epsilon. \tag{58}$$

As before, the "easy" sampling and search problems are

$$\mathsf{SampQNC}^0 \coloneqq \bigcap_{\epsilon>0} \mathsf{SampQNC}^0_\epsilon \quad \text{and} \quad \mathsf{FQNC}^0 \coloneqq \bigcap_{\epsilon\in(0,\frac{1}{2})} \mathsf{FQNC}^0_\epsilon. \tag{59}$$

Since it is possible to simulate the standard classical basis $B_0 = \{\wedge_2, \vee_2, \neg\}$ with Clifford+T gates [55], quantum shallow circuits are at least as powerful as classical shallow circuits. We therefore have the trivial inclusions:

$$\mathsf{SampNC}^0/\mathsf{rpoly} \subseteq \mathsf{SampQNC}^0 \quad \text{and} \quad \mathsf{FNC}^0/\mathsf{rpoly} \subseteq \mathsf{FQNC}^0. \tag{60}$$

In other words, among the "easy" sampling and search problems that quantum shallow circuits can solve are the "easy" sampling and search problems that classical shallow circuits can solve. The same is manifestly true for the "hard" problems: for any valid $\epsilon > 0$ it holds that quantum shallow circuits are at least as powerful as classical shallow circuits:

$$\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon \subseteq \mathsf{SampQNC}^0_\epsilon \quad \text{and} \quad \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon \subseteq \mathsf{FQNC}^0_\epsilon. \tag{61}$$

We will interrogate these relationships in more detail in Sec. 3.5.

Given a quantum shallow circuit $\mathcal{Q}_n$, the input $|0\rangle^{\otimes n}$ generates an output state $\rho^{\mathcal{Q}}_n$. If we measure this state in different bases, then the output distributions we obtain will in most cases be different. Such an idea is the germ of an alternative quantum computational model originally developed by Raussendorf, Browne, and Briegel: *measurement-based quantum computation* (MBQC) [62, 63].

Roughly speaking, an MBQC computer, also called a *one-way quantum computer*, works by first preparing a highly entangled state known as a *cluster state* and then adaptively measuring the qubits in a way that solves the computational problem. Thus, the input to a one-way quantum computer encodes the basis and order in which the individual qubits in the cluster are measured [64]. In a way, this introduces a temporal order to the computation, so it is ostensibly well-suited to a qDAG-like representation [63]. MBQC is plainly contra to quantum circuits, where a computational problem is encoded as a product state, acted on by a series of Clifford+T gates, and then measured in the computational basis. In a quantum circuit, there is no clear temporal order of events, and so the quantum circuit model is generally ill-suited to a qDAG-like representation.

As mentioned, the main resource used in MBQC is a *cluster state*. This is an instance of a more general state known as a *graph state*. Given an undirected graph $G = (V, E)$, the *graph state* for G is defined as

$$|G\rangle \coloneqq \prod_{e \in E} \mathrm{CZ}_e \bigotimes_{v \in V} |+\rangle_v, \tag{62}$$

where CZ_e is the controlled-Z operation acting on the registers in the edge e and $|+\rangle_v = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. A *$k$-dimensional cluster state* $|G\rangle$ is a graph state in which G is a k-dimensional lattice.

The notational similarity between Eqs. 53 and 62 is manifest. It makes it sensible to wonder, what depth quantum circuit is required to prepare a graph state $|G\rangle$ on $|V| = n$ vertices?

Starting in the state $|0\rangle^{\otimes n}$, one can prepare $\otimes_{v\in V}|+\rangle_v$ in constant depth with the Clifford+T gate set because $\otimes_{v\in V}|+\rangle_v = H^{\otimes n}|0\rangle^{\otimes n}$. It remains to modify the phases between the qubits as specified by the edges of the graph G. By Eq. (62), we accomplish this by applying a CZ_e gate to each edge $e \in E$. Therefore, the size of the circuit is at most $n + |E|$ and the depth is at most $1 + |E|$. By the handshaking lemma (1), $|E| = O(n)$. Therefore, graph states can be prepared on quantum circuits with at most linear size and linear depth.

Recalling the discussion of undirected graphs in Sec. 1.1, it is not difficult to see that the depth can be reduced through parallelization and an edge coloring $c : E \to \{1, \ldots, \chi_c\}$. Since for all edge colorings c it holds that $\chi_c \leq \Delta(G) + 1$, it follows that the quantum circuit needs at most $\Delta(G) + 1$ layers to implement the |E| many CZ_e gates. Hence, the depth required to make a graph state is $O(\Delta(G))$ [65]. Since $\Delta(G)$ is constant for any k-dimensional lattice graph G, a k-dimensional cluster state can be prepared on a constant depth quantum circuit [66]. This conclusion is interesting, because dimension two or higher cluster states constitute a *universal resource* for MBQC, meaning quantum circuits and one-way quantum computers can simulate each other [62, 63, 67, 68].

In particular, a one-way quantum computer can simulate a quantum shallow circuit $\mathcal{Q}$. This implies that given a graph state $|G\rangle$ and an input $\langle x\rangle$, there exists a completely-positive and trace-preserving map $\mathcal{P}(x)$ consisting of CZ operations, measurements M^α in the basis $\frac{1}{\sqrt{2}}(|0\rangle \pm e^{i\alpha}|1\rangle)$, where $\alpha \in [0, 2\pi)$, and X and Z Pauli operations such that $\mathcal{P}(x)|G\rangle\langle G| = \mathcal{Q}|x\rangle\langle x|\mathcal{Q}^\dagger$ [69]. In MBQC, the map $\mathcal{P}$ is called a *measurement pattern* and is evidently the MBQC equivalent of a quantum circuit. It follows, therefore, that given a family of quantum shallow circuits $(\mathcal{Q}_n)_{n\in\mathbb{N}}$, there exists a family of cluster states and measurement patterns $(\rho_n, \mathcal{P}_n)_{n\in\mathbb{N}}$ such that for all $x \in \{0,1\}^*$ it holds that $\mathcal{Q}_n|x\rangle\langle x|\mathcal{Q}^\dagger = \mathcal{P}_{|x|}(x)\rho_{|x|}$.[14]

Of course, in order to fairly compare a one-way quantum computer to shallow circuits, we need an understanding of how the complexity measures in each relate. Whereas the formal details of this comparison require an exposition of the *measurement calculus*, which is essentially an algebraic framework that neatly characterizes MBQC [69], it suffices for our purposes to provide a rough outline of the ideas. The main references are [63] and [69].

Similar to quantum circuits, there are two complexity measures that are

[14] Here the cluster state $\rho_{|x|}$ at most depends on the size of x but not x itself.

distinguished when talking about measurement patterns. Coincidentally, they are also called *size* and *depth*. The *size* of a measurement pattern $\mathcal{P}$ is the number of CZ operations, M^α measurements, and X and Z Pauli operations that constitute $\mathcal{P}$. On the other hand, the *depth* of $\mathcal{P}$ is, roughly speaking, the longest subsequence of operators in $\mathcal{P}$ that "depend" on the output of an operator that precedes it [70]. Such dependency has to do with the adaptive nature of the computation; it is made explicit in the measurement calculus formalism [69].

With these definitions, the complexity of a quantum computer can be related to the complexity of the one-way computer that simulates it [63]:

Proposition 16. *If $\mathcal{Q}$ is a quantum circuit of size $s_\mathcal{Q}$ and depth $d_\mathcal{Q}$, then $\mathcal{Q}$ can be simulated by a measurement pattern $\mathcal{P}$ of size $O(s_\mathcal{Q})$ and depth $O(d_\mathcal{Q})$.*

Consequently, a family of quantum shallow circuits $(\mathcal{Q}_n)_{n\in\mathbb{N}}$ can be simulated by a family of cluster states and *constant depth* measurement patterns $(\rho_n, \mathcal{P}_n)_{n\in\mathbb{N}}$.

Based on the information so far, it is natural to conjecture that there is a complexity theoretic equivalence here—namely, that what can be computed in constant depth and polynomial size on a quantum circuit can also be computed in constant depth and polynomial size on a one-way quantum computer. However, this is false because no constant depth quantum circuit can determine if an input $\langle x\rangle$ has an even or odd number of bits equal to 1,[15] whereas a constant depth one-way computer can [71]. Consequently, the simulation of a one-way computer on a quantum circuit requires a deeper depth. How much deeper was established by Broadbent and Kashefi [70, 72]:

Proposition 17. *If $\mathcal{P}$ is a measurement pattern of size $s_\mathcal{P}$ and depth $d_\mathcal{P}$, then the action of $\mathcal{P}$ on any k-dimensional cluster state can be simulated by a quantum circuit $\mathcal{Q}$ of size $O(s_\mathcal{P}^3)$ and depth $O(d_\mathcal{P} \log s_\mathcal{P})$.*

Consequently, if $\mathcal{P}$ is a *constant size* and constant depth measurement pattern, then it can be simulated on a constant size quantum shallow circuit. However, if $\mathcal{P}$ is *polynomial size* and constant depth, then it cannot in general be simulated on a polynomial size quantum shallow circuit. Interestingly, if we allow quantum circuits to have unbounded fan-out gates, then the depth complexities become the same [70].

[15] This is a famous decision problem in complexity theory called *parity*.

Let $\mathcal{P}$ be a measurement pattern with polynomial size $s_{\mathcal{P}}$ and unit depth $d_{\mathcal{P}} = 1$. By Proposition 17, such a pattern is not necessarily simulable by a quantum shallow circuit. However, the unit depth restriction implies that each operation in $\mathcal{P}$ acts on distinct qubits in the cluster state. Hence, a unit depth measurement pattern acts on single qubits only, and hence can be expressed as the tensor product

$$\mathcal{P} = \bigotimes_{i=1}^{s_{\mathcal{P}}} A_i, \tag{63}$$

where, in general, each A_i is a sequence of measurements M^{α_i} and X_i and Z_i Pauli operations that can depend on the input to the computer.

It is proved in [70] that to simulate the X_i and Z_i Pauli operations in a unit depth measurement pattern requires a shallow circuit with unbounded fan-out. As we want bounded fan-out, we restrict the measurement pattern $\mathcal{P}$ to be *measure-only*, that is, we require that each $A_i = M^{\alpha_i}$ for some $\alpha_i \in [0, 2\pi)$. It is easily verified that $M^{\alpha_i} = HR_z(-\alpha_i)$. Consequently, provided α_i is such that $R_z(-\alpha_i)$ can be computed in constant depth, the polynomial size and constant depth measurement pattern (63) can be simulated on a polynomial size quantum shallow circuit. In particular, there exists a quantum shallow circuit $\mathcal{Q}$ such that, for some cluster state ρ, the input $\langle x\rangle$ implies

$$\Pr_{\mathtt{Y}\sim\mathscr{D}(\mathcal{Q})}(\mathtt{Y}) = \mathrm{Tr}(\rho|\mathtt{Y}\rangle\langle\mathtt{Y}|\mathcal{P}(x)). \tag{64}$$

As $|\mathtt{Y}\rangle\langle\mathtt{Y}| = |\mathtt{Y}_1\rangle\langle\mathtt{Y}_1| \otimes \cdots \otimes |\mathtt{Y}_{|x|}\rangle\langle\mathtt{Y}_{|x|}|$ is a product of single qubit projectors, the product $|\mathtt{Y}_i\rangle\langle\mathtt{Y}_i|M^{\alpha_i}(x) = M^{\alpha_i}_{\mathtt{Y}_i}(x)$. Furthermore, like we did for classical and quantum shallow circuits, we can imagine that the input $\langle x\rangle$ is a valuation of some random variable $\mathtt{X}$. Then, the circuit that simulates the measure-only measurement pattern $\mathcal{P}$ satisfies, for some cluster state ρ, the conditional probability distribution

$$\Pr_{\mathtt{Y}|\mathtt{X}\sim\mathscr{D}(\mathcal{Q})}(\mathtt{Y} \mid \mathtt{X}) = \mathrm{Tr}\left(\rho M^{\alpha_1}_{\mathtt{Y}_1}(\mathtt{X}) \otimes \cdots \otimes M^{\alpha_{|\mathtt{X}|}}_{\mathtt{Y}_{|\mathtt{X}|}}(\mathtt{X})\right). \tag{65}$$

It is now evident that this distribution is compatible with $\mathrm{QGHZ}_{|x|,|x|+1}$. In particular, if $M^{\alpha_i}_{\mathtt{Y}_i}$ only depends on at most m input bits $\mathtt{X}_{i_1}, \ldots, \mathtt{X}_{i_m}$, then the distribution is compatible with $\mathrm{QGHZ}_{n,m}$.

Theorem 18. *If $\mathcal{P}_n$ is a measure-only measurement pattern on an n qubit cluster state such that the measurement operators can be realized in constant*

depth, then there exists a quantum shallow circuit $\mathcal{Q}_n$ *that can simulate* $\mathcal{P}_n$ *and whose output distribution* $\mathscr{D}(\mathcal{Q}_n)$ *is compatible with* $\mathrm{QGHZ}_{n,m}$ *for some* $m \geq 0$.

3.5 Quantum Advantage and Discussion

By "quantum advantage" ("quantum speedup", "quantum supremacy", etc.) one typically means that there exists a computational task that a quantum computer can efficiently solve that no classical computer can efficiently solve. Hence, by this definition, if there is a quantum advantage, then the extended Church-Turing thesis (ECT) is false. Indeed, this type of quantum advantage is what is at stake by the conjectured inequivalence of BPP and BQP, where BQP is the quantum analogue of BPP [44]. However, there is a weaker notion of quantum advantage that does not entail violating the ECT. This second notion is the one we mean in this essay.

More precisely, we are concerned with a potential advantage with quantum shallow circuits. Given the two main types of complexity classes we have defined ("easy" and "hard" sampling/search problems), there are naturally two different types of advantage one can speak of.

Definition 9 (Weak/Strong Sampling Advantage)**.** We say quantum shallow circuits offer a *weak sampling advantage* if and only if there exists $\epsilon > 0$ such that

$$\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon \subsetneq \mathsf{SampQNC}^0_\epsilon. \tag{66}$$

On the other hand, quantum shallow circuits offer a *strong sampling advantage* if and only if there exists $\epsilon > 0$ such that

$$\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon \subsetneq \mathsf{SampQNC}^0. \tag{67}$$

Thus, there is a strong sampling advantage if and only if there is a sampling problem that is "easy" for some quantum shallow circuit but "hard" for every classical shallow circuit. The definitions of *weak* and *strong search advantage* are analogous.

Evidently, a strong sampling (search) advantage implies a weak sampling (search) advantage. What is more interesting, though, is that a strong sampling advantage implies a strong search advantage. This more or less follows because search problems are instances of sampling problems.

Lemma 19. $\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon = \mathsf{SampQNC}^0 \implies \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon = \mathsf{FQNC}^0$.
(Proof.)

Likewise, a weak sampling advantage implies a weak search advantage by essentially the same logic.

Lemma 20. $\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon = \mathsf{SampQNC}^0_\epsilon \implies \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon = \mathsf{FQNC}^0_\epsilon$.
(Proof.)

In 2018, Bravyi, Gosset, and König proved a strong search advantage for quantum shallow circuits [11]. By Theorem 19, there is thus a strong search advantage for quantum shallow circuits as well. This was quickly extended to more powerful complexity classes [61]. A natural question is if this advantage is robust to noise. Indeed, it is shown in [54] and [73] that the advantage of quantum shallow circuits is noise tolerant in the sense that the advantage with noise is weak (though hardly, $\epsilon < 0.01$ [54]). That a strong advantage without noise will generically lead to no advantage or a weak advantage is easy to see with a primitive constant error model, at least for sampling problems.

Let $S = (\mathscr{D}_x)_{x\in\{0,1\}^*} \in \mathsf{SampQNC}^0 \backslash \mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon$ be a sampling problem for which there is a strong sampling advantage (in [11], $\epsilon < 1/8$ suffices). Then by definition there exists a family of quantum shallow circuits $(\mathcal{Q})_{n\in\mathbb{N}}$ and a family of classical shallow circuits $(\mathcal{C}_{n+r})_{n\in\mathbb{N}}$ such that the input $\langle x\rangle$ implies

$$\forall \delta > 0 : \left\| \mathscr{D}(\mathcal{Q}_{|x|}) - \mathscr{D}_x \right\| \leq \delta \quad \text{and} \quad \left\| \mathscr{D}(\mathcal{C}_{|x|}) - \mathscr{D}_x \right\| \leq \epsilon. \tag{68}$$

But with constant noise, each $\mathscr{D}(\mathcal{Q}_{|x|})$ is instead some other distribution $\widetilde{\mathscr{D}}(\mathcal{Q}_{|x|})$ satisfying $\left\| \mathscr{D}(\mathcal{Q}_{|x|}) - \widetilde{\mathscr{D}}(\mathcal{Q}_{|x|}) \right\| \leq \eta_{|x|}$, where $\eta_{|x|}$ is the noise rate on $|x|$ qubits. By the triangle inequality,

$$\begin{aligned} \forall \delta > 0 : \left\| \widetilde{\mathscr{D}}(\mathcal{Q}_{|x|}) - \mathscr{D}_x \right\| &= \left\| \widetilde{\mathscr{D}}_{\mathcal{Q}_{|x|}} - \mathscr{D}(\mathcal{Q}_{|x|}) + \mathscr{D}(\mathcal{Q}_{|x|}) - \mathscr{D}_x \right\| && (69) \\ &\leq \left\| \widetilde{\mathscr{D}}(\mathcal{Q}_{|x|}) - \mathscr{D}(\mathcal{Q}_{|x|}) \right\| + \left\| \mathscr{D}(\mathcal{Q}_{|x|}) - \mathscr{D}_x \right\| && (70) \\ &\leq \eta_{|x|} + \delta. && (71) \end{aligned}$$

Consequently, if it holds that $\eta \coloneqq \limsup_{n\to\infty} \eta_n < \epsilon$, then the most we can conclude is that $S \in \mathsf{SampQNC}^0_\eta \backslash \mathsf{SampNC}^0/\mathsf{rpoly}_\eta$—that is, there exists a weak sampling advantage. As far as we know, it is an open question whether

the strong sampling advantage of quantum shallow circuits is robust, let alone their strong search advantage.

In [17], Aaronson proves the "sampling and search equivalence theorem", which ultimately entails that $\mathsf{SampBPP} = \mathsf{SampBQP}$ if and only if $\mathsf{FBPP} = \mathsf{FBQP}$. In other words, efficient classical computers and efficient quantum computers sample from the same distributions if and only if efficient classical computers and efficient quantum computers solve the same search problems. This is fascinating, because it implies that a quantum advantage at the level of sampling implies a quantum advantage at the level of search.

Whereas the forward direction of the sampling/search equivalence theorem holds for shallow circuits (Lemmas 19 and 20), attempting to apply Aaronson's argument to the other direction seems to fail. The main impediment is that a classical shallow circuit can most likely not choose an input bit uniformly at random. We say "most likely" because we have not found a proof of this fact, but it seems plausible based on the limited computational abilities of shallow circuits.

While there are also some more subtle details up in the air, we do propose the following conjectures that are still very much in the spirit of Aaronson's sampling/search equivalence theorem. The idea is simple: since the main impediment is an inability to uniformly choose an input bit, simply empower the circuits with an oracle $\mathcal{U}$ that does it for us.

Conjecture 21. $(\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon)^{\mathcal{U}} = (\mathsf{SampQNC}^0)^{\mathcal{U}} \iff (\mathsf{FNC}^0/\mathsf{rpoly}_\epsilon)^{\mathcal{U}} = (\mathsf{FQNC}^0)^{\mathcal{U}}$.

Similarly, we expect this to hold even at the level of a weak advantage.

Conjecture 22. $(\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon)^{\mathcal{U}} = (\mathsf{SampQNC}^0_\epsilon)^{\mathcal{U}} \iff (\mathsf{FNC}^0/\mathsf{rpoly}_\epsilon)^{\mathcal{U}} = (\mathsf{FQNC}^0_\epsilon)^{\mathcal{U}}$.

If it could be shown that $\mathcal{U}$ or something equivalent to it can be simulated on shallow circuits, then it would suffice to prove a strong sampling advantage to prove a strong search advantage. By Lemma 20, the simple robustification argument given above would then carry through to search problems as well.

Although we have not been able to prove a weak sampling or search advantage with quantum shallow circuits using the tools developed herein (let alone a strong sampling or search advantage), our results point toward what we find to be an intriguing way forward. By Theorems 8 and 18, if

$\mathcal{C}_{n+r}$ is a classical shallow circuit of depth d and fan-in k and $\mathcal{Q}_n$ is quantum shallow circuit over the Clifford+T gate set representing a measure-only measurement pattern $\mathcal{P}$, then $\mathscr{D}(\mathcal{C}_{n+r})$ is compatible with GHZ_{n,k^d} and $\mathscr{D}(\mathcal{Q}_n)$ is compatible with some $\mathrm{QGHZ}_{n,m}$. Therefore, if $\mathcal{Q}_n$ is such that $\mathscr{D}(\mathcal{Q}_n) \in \mathscr{P}(\mathrm{QGHZ}_{n,m})\backslash\mathscr{P}(\mathrm{GHZ}_{n,k^d})$—that is, if $\mathscr{D}(\mathcal{Q}_n)$ is in the QC-gap of GHZ_{n,k^d} relative to $\mathrm{QGHZ}_{n,m}$—then it follows that no classical shallow circuit can sample from $\mathscr{D}(\mathcal{Q}_n)$ exactly.

Of course, provided $n > k^d + 1$, the QC-gap is non-empty (Theorem 6). But the question is if there exists a distribution in the QC-gap that can be realized by a quantum shallow circuit. In other words, is there a measure-only measurement pattern $\mathcal{P}$ on n qubits such that its output distribution is in the QC-gap? One potential way forward is to use the fact that to every graph state $|G\rangle$ there corresponds a generalized Bell inequality $\mathcal{B}_G$ [65, 74–76]. Thus, for a family of cluster states $(\rho_n)_{n\in\mathbb{N}}$, if it could be shown that their corresponding generalized Bell inequalities $\mathcal{B}_{\rho_n}$ as described in [75] can be realized by a family of measure-only measurement pattern $(\mathcal{P}_n)_{n\in\mathbb{N}}$, and thus by a family of quantum shallow circuits $(\mathcal{Q}_n)_{n\in\mathbb{N}}$, then the *Bell sampling problem*

$$S_{\mathcal{B}} \coloneqq \Big(\mathscr{D}_x \mid \mathcal{B}_{\rho_{|x|}}(\boldsymbol{p}) > 0\Big)_{x\in\{0,1\}^*}, \tag{72}$$

where $\boldsymbol{p} \in \Delta_{2^{|x|}}$ is the point in the probability simplex $\Delta_{2^{|x|}}$ corresponding to $\mathscr{D}_x$, would not only witness the QC-gap, but would also prove a strong sampling advantage. This would, in other words, demonstrate a quantum advantage based purely on a QC-gap. This research is ongoing and is something on which we hope to report soon.

Acknowledgements

This year at Perimeter Institute has been the most intellectually stimulating of the ones I've experienced so far. I thank all the people who have made it special, only a proper subset of which I include here.

Thank you to Rob Spekkens for agreeing to be my advisor, for making me think harder than ever before about my normative assumptions of the world, and for proposing this project. I have learned a lot from it and even more from you.

Thank you to Marina Maciel Ansanelli, T.C. Fraser, Elie Wolfe, and Rob Spekkens (part of the DAG winter school group) for introducing me to the wonder theory that is causal inference. It has given me a lifetime of questions to think about.

Thank you to Thomas Galley and Flaminia Giacomini for showing an interest in my ideas related to semi-classical gravity. I hope they turn out to be fruitful soon.

Thank you to Philippe Allard Guerin for being a great T.A. this past year and for entertaining my questions and ideas about quantum foundations and Turing machines.

Thank you to the whole PSI cohort for the memories we've shared and for making this year as enjoyable as it was.

Thank you to Sotiris μαλάκας Mygdalas for the bond we've formed, the laughs we've had, and the Schuller we've enjoyed. You kept me sane during those courses with which I had a most poor relationship.

Thank you to Charles Cummings for all the things you've taught me and the philosophy we've explored.

Thank you to Sophia Gonzalez Garcia for the many memories we've shared and the DAGs we've conditioned.

Thank you to Manu Srivastava for the type-II weekend struggles we've endured and the DAGs we've marginalized.

Thank you to Gabriel Golfetti for listening to my many crazy ideas and for never failing to improve on them.

Thank you to Anna Knörr for showing me what curiosity actually looks like.

Thank you to Eivind Jørstad for the adventures we've had and the climbing stoke you carry.

Thank you to Mathew One Bub for the philosophizing we've done and

the conclusions we've failed to reach.

Thank you to Jaime Redondo Yuste for the heart you carry and the intellect you wield.

Thank you to Jordan Krywonos for your inspiring work ethic and the laughs we've shared.

Thank you to Javi Hernández Morales for never failing to bring a smile to my face.

Thank you to Amirreza Negari for your kind heart and the memories we've shared.

Thank you to Dan Sehayek for the friendship we've formed and the moments we'll soon share at UCSB.

Thank you to Tomáš Vávra and Maegan-Rose Schwarz for making a year of living in Waterloo and climbing inside bearable.

Thank you to Chai Karamchedu and Daniel Bashir for exercising my mind outside of physics this year and teaching me oh so much.

Thank you to Adam Bentson, Chuck Hendrick, and Jaine Mueller for starting me on this path in life. Thank you to Brian Shuve, Tom Donnelly, Nicholas Breznay, Theresa Lynn, and Jason Gallicchio for encouraging me to stay on it.

And thank you to my family—Noni, Debra, Jake, Ben, Emma, Steve, and Linda—for your unconditional love and support. I would not have the life I do without you.

References

[1] R. P. Feynman, "Simulating physics with computers," *Int. J. Theor. Phys.* **21** (1982) 467.

[2] K. Gödel, "Über formal unentscheidbare sätze der Principia Mathematica und verwandter systeme I," *Monatshefte für Mathematik und Physik* **38** (1931) 173.

[3] G. Etesi and I. Németi, "Non-Turing computations in Malament-Hogarth spacetimes," *Int. J. Theor. Phys.* **41** (2002) 341.

[4] D. Deutsch, "Quantum mechanics near closed timelike lines," *Phys. Rev. D* **44** (1991) 3197.

[5] S. Aaronson and J. Watrous, "Closed timelike curves make quantum and classical computing equivalent," *Proc. R. Soc. A.* **465** (2009) 631.

[6] L. K. Grover, "Quantum mechanics helps in searching for a needle in a haystack," *Phys. Rev. Lett.* **79** (1997) 325.

[7] P. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," *Proc. 35th Annual Symposium on Foundations of Computer Science* (1994) 124.

[8] D. Gottesman, "The heisenberg representation of quantum computers," *Proc. of the XXII International Colloquium on Group Theoretical Methods in Physics* (1999) 32.

[9] M. Howard *et al.*, "Contextuality supplies the magic for quantum computation," *Nature* **510** (2014) 351.

[10] H. Buhrman *et al.*, "Quantum communication complexity advantage implies violation of a Bell inequality," *PNAS* **113** (2016) 3191.

[11] S. Bravyi, D. Gosset, and R. König, "Quantum advantage with shallow circuits," *Science* **362** (2018) 308.

[12] J. Henson, R. Lal, and M. F. Pusey, "Theory independent limits on correlations from generalized Bayesian networks," *New J. Phys.* **16** (2014) 113043.

[13] C. J. Wood and R. W. Spekkens, "The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell inequality violations require fine-tuning," *New J. Phys.* **17** (2015) 033002.

[14] T. Fritz, "Beyond Bell's theorem: correlation scenarios," *New J. Phys.* **14** (2012) 103001.

[15] S. G. Garcia *et al.* In preparation.

[16] R. Diestel, *Graph Theory.* Springer, New York, 2005.

[17] S. Aaronson, "The equivalence of sampling and searching," *Theory Comput. Syst.* **55** (2014) 281.

[18] A. B. Watts, R. Kothari, L. Schaeffer, and A. Tal, "Exponential separation between shallow quantum circuits and unbounded fan-in shallow classical circuits," *Proc. of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (2019) 515.

[19] J. Pearl, *Causality.* Cambridge University Press, Cambridge, 2009.

[20] R. J. Evans, "Graphs for margins of Bayesian networks," *Scand. J. Statist.* **43** (2016) 625.

[21] R. J. Evans, "Margins of discrete Bayesian networks," *Ann. Statist.* **46** (2018) 2623.

[22] L. D. Garcia, M. Stillman, and B. Sturmfels, "Algebraic geometry of Bayesian networks," *J. Symb. Comput.* **39** (2005) 331.

[23] P. Zwiernik, *Semialgebraic Statistics and Latent Tree Models.* CRC Press, Boca Raton, 2016.

[24] M. Studený, *Probabilistic Conditional Independence Structures.* Springer, London, 2004.

[25] M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, *Algebraic Aspects of Conditional Independence and Graphical Models*, ch. 3. CRC Press, 2018.

[26] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, New York, 2007.

[27] R. W. Spekkens, "Evidence for the epistemic view of quantum states: a toy theory," *Phys. Rev. A* **75** (2007) 032110.

[28] S. D. Bartlett, T. Rudolph, and R. W. Spekkens, "Reconstruction of Gaussian quantum mechanics from Liouville mechanics with an epistemic restriction," *Phys. Rev. A* **86** (2012) 012103.

[29] R. G. Winther, "The structure of scientific theories," *The Stanford Encyclopedia of Philosophy* (2021) .

[30] N. Harrigan and R. Spekkens, "Einstein, incompleteness, and the epistemic view of quantum states," *Found. Phys.* **40** (2010) 125.

[31] J. Barrett, R. Lorenz, and O. Oreshkov, "Quantum causal models," `arXiv:1906.10726 [quant-ph]`.

[32] J. A. Allen *et al.*, "Quantum common causes and quantum causal models," *Phys. Rev. X* **7** (2017) 031021.

[33] J. Bell, "On the Einstein Podolsky Rosen paradox," *Phys. Phys. Fiz.* **1** (1964) 195.

[34] J. Bell, *Speakable and Unspeakable in Quantum Mechanics*. Cambridge University Press, Cambridge, 2004.

[35] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt, "Proposed experiment to test local hidden-variable theories," *Phys. Rev. Lett.* **23** (1970) 880.

[36] T. C. Fraser and E. Wolfe, "Causal compatibility inequalities admitting quantum violations in the triangle structure," *Phys. Rev. A* **98** (2018) 022113.

[37] J. Pienaar, "Which causal structures might support a quantum-classical gap?" *New J. Phys.* **19** (2017) 043021.

[38] D. Greenberger, M. Horne, A. Shimony, and A. Zeilinger, "Bell's theorem without inequalities," *Am. J. Phys.* **58** (1990) 1131.

[39] D. Mermin, "Quantum mysteries revisited," *Am. J. Phys.* **58** (1990) 731.

[40] B. F. Toner and D. Bacon, "The communication cost of simulating Bell correlations," *Phys. Rev. Lett.* **91** (2003) 187904.

[41] T. E. Tessier, I. H. Deutsch, and C. M. Caves, "Efficient classical-communication-assisted local simulation of n-qubit GHZ correlations," `arXiv:quant-ph/0407133 [quant-ph]`.

[42] A. M. Turing, "On computable numbers, with an application to the Entscheidungproblem," *Proc. Lond. Math. Soc.* **42** (1936) 230.

[43] A. M. Turing, "Intelligent machinery," *National Physical Laboratory* (1948) 3.

[44] S. Aaronson, *Quantum Computing Since Democritus*. Cambridge University Press, Cambridge, 2013.

[45] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*. Cambridge University Press, Cambridge, 2007.

[46] J. B. Copeland, "The Church-Turing thesis," *The Stanford Encyclopedia of Philosophy* (2020) .

[47] O. Goldreich, *Computational Complexity: A Conceptual Perspective*. Cambridge University Press, Cambridge, 2008.

[48] L. Fortnow, "The status of the P versus NP problem," *Commun. of the ACM* **52** (2009) 78.

[49] H. Vollmer, *Introduction to Circuit Complexity: A Uniform Approach*. Springer-Verlag, Berlin, 1999.

[50] L. M. Adelman, "Two theorems on random polynomial time," *IEEE* (1978) 75.

[51] N. Pippenger, "On simultaneous resource bounds," *IEEE* (1979) 307.

[52] S. Cook, "A taxonomy of problems with fast parallel algorithms," *Inf. Control.* **64** (1985) 2.

[53] M. Ajtai and M. Ben-Or, "A theorem on probabilistic constant depth computations," *Proc. of the Sixteenth Annual ACM Symposium on Theory of Computing* (1984) 471.

[54] S. Bravyi, D. Gosset, R. König, and M. Tomamichel, "Quantum advantage with noisy shallow circuits in 3D," *Nat. Phys.* **16** (2020) 1040.

[55] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, 2000.

[56] C. A. Pèrez-Delgado and P. Kok, "Quantum computers: definition and implementations," *Phys. Rev. A* **83** (2011) 012303.

[57] D. Deutsch, "Quantum theory, the Church-Turing principle, and the universal quantum computer," *Proc. R. Soc. Lond. A* **400** (2018) 97.

[58] E. Farhi, J. Goldstone, S. Gutmann, and M. Sipser, "Quantum computations by adiabatic evolution," `arXiv:quant-ph/0001106 [quant-ph]`.

[59] A. C. Yao, "Quantum circuit complexity," *IEEE* (1993) 352.

[60] V. Scarani, A. Acìn, E. Schenck, and M. Aspelmeyer, "Nonlocality of cluster states of qubits," *Phys. Rev. A* **71** (2005) 042325.

[61] D. Grier and L. Schaeffer, "Interactive shallow Clifford circuits: quantum advantage against NC^1 and beyond," *Proc. of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (2020) 875.

[62] R. Raussendorf and H. J. Briegel, "A one-way quantum computer," *Phys. Rev. Lett.* **86** (2001) 5188.

[63] R. Raussendorf, D. E. Browne, and H. J. Briegel, "Measurement-based quantum computation on cluster states," *Phys. Rev. A* **68** (2003) 022312.

[64] B. M. Terhal and D. P. DiVincenzo, "Adaptive quantum computation, constant depth quantum circuits and Arthur-Merlin games," *Quant. Inf. Comp.* **4** (2004) 134.

[65] A. Cabello *et al.*, "Nonlocality for graph states," *Laser Phys.* **18** (2008) 335.

[66] P. Aliferis and D. W. Leung, "Computation by measurements: a unifying picture," *Phys. Rev. A* **70** (2004) 062314.

[67] M. A. Nielsen, "Cluster-state quantum computation," *Rep. Math. Phys.* **57** (2006) 147.

[68] M. Hein *et al.*, "Entanglement in graph states and its applications," *Proc. of the International School of Physics "Enrico Fermi"* **162** (2006) 115.

[69] V. Danos, E. Kashefi, and P. Panangaden, "The measurement calculus," *J. ACM* **54** (2007) 8.

[70] D. Browne, E. Kashefi, and S. Perdix, "Computational depth complexity of measurement-based quantum computation," `arXiv:0909.4673 [quant-ph]`.

[71] M. Fang *et al.*, "Quantum lower bounds for fanout," *Quantum Inf. Comput.* **6** (2006) 46.

[72] A. Broadbent and E. Kashefi, "Parallelizing quantum circuits," *Theor. Comput. Sci.* **410** (2009) 2489.

[73] R. Mezher *et al.*, "Fault-tolerant quantum speedup from constant depth quantum circuits," *Phys. Rev. Research* **2** (2020) 033444.

[74] A. Cabello *et al.*, "Optimal preparation of graph states," *Phys. Rev. A* **83** (2011) 042314.

[75] O. Gühne *et al.*, "Bell inequalities for graph states," *Phys. Rev. Lett.* **95** (2005) 120405.

[76] M. Niu *et al.*, "A note of Bell inequalities for graph states," *Int. J. Theor. Phys.* **60** (2021) 2511.

A Two Tricks for Proving a QC-Gap

Let $\mathtt{mDAG}$ and $\mathtt{qDAG}$ denote the set of all mDAGs and qDAGs, respectively. We say a map $\mathfrak{X} : \mathtt{mDAG} \times \mathtt{qDAG} \to \mathtt{mDAG} \times \mathtt{qDAG}$ is *QC-gap non-increasing for* $G \in \mathtt{mDAG}$ *and* $QG \in \mathtt{qDAG}$ if and only if

$$(G, QG) \overset{\mathfrak{X}}{\longmapsto} (G^{\mathfrak{X}}, QG^{\mathfrak{X}}) \implies \mathscr{P}(QG) \backslash \mathscr{P}(G) \supseteq \mathscr{P}(QG^{\mathfrak{X}}) \backslash \mathscr{P}(G^{\mathfrak{X}}). \quad (73)$$

While we prefer to characterize QC-gap non-increasing maps in terms of compatible distributions, there is an equivalent characterization in terms of statistical models (which have the intuitive advantage of being geometrical in nature). Namely, Eq. (73) holds, and hence $\mathfrak{X}$ is QC-gap non-increasing, if and only if

$$(G, QG) \overset{\mathfrak{X}}{\longmapsto} (G^{\mathfrak{X}}, QG^{\mathfrak{X}}) \implies \mathcal{M}(QG) \backslash \mathcal{M}(G) \supseteq \mathcal{M}(QG^{\mathfrak{X}}) \backslash \mathcal{M}(G^{\mathfrak{X}}). \quad (74)$$

Though trivial, the following theorem makes it plain that the terminology "QC-gap non-increasing" is sensible:

Theorem 23. *If $\mathfrak{X}$ is QC-gap non-increasing for G and QG, then G admits a QC-gap relative to QG if $G^{\mathfrak{X}}$ admits a QC-gap relative to $QG^{\mathfrak{X}}$.*

In [15], the authors call QC-gap non-increasing maps *tricks*. Below we describe two such tricks for the mDAG $\mathrm{GHZ}_{n,m}$ and the qDAG QGHZ_n. The general case is obviously more complicated.

Recall that $\mathrm{GHZ}_{n,m}$ is the mDAG $(V \cup \{\Lambda\}, E)$ where the visible vertices $V = \{\mathtt{X}_1, \ldots, \mathtt{X}_n, \mathtt{Y}_1, \ldots, \mathtt{Y}_n\}$, for all $\mathtt{X}_i \in V$ it holds that $\mathtt{pa}(\mathtt{X}_i) = \varnothing$, and for all $\mathtt{Y}_i \in V$ it holds that $\mathtt{lpa}(\mathtt{Y}_i) = \{\Lambda\}$ and $|\mathtt{vpa}(\mathtt{Y}_i)| = m + 1$. QGHZ_n, on the other hand, is isomorphic to $\mathrm{GHZ}_n = \mathrm{GHZ}_{n,0}$ as a directed graph.

Definition 10 (GHZ Conditioning Trick)**.** Let $\mathtt{X}$ be a visible vertex in $\mathrm{GHZ}_{n,m}$ satisfying $\mathtt{pa}(\mathtt{X}) = \varnothing$. The *GHZ conditioning trick* is the map

$$\mathrm{cond}_{\mathtt{X}} : (\mathrm{GHZ}_{n,m}, \mathrm{QGHZ}_n) \mapsto (\mathrm{GHZ}_{n,m}^{\mathrm{cond}_{\mathtt{X}}}, \mathrm{QGHZ}_n^{\mathrm{cond}_{\mathtt{X}}}), \quad (75)$$

where $\mathrm{QGHZ}_n^{\mathrm{cond}_{\mathtt{X}}}$ is QGHZ_n but with $\mathtt{X}$ removed (as well as the edge stemming from it) and $\mathrm{GHZ}_{n,m}^{\mathrm{cond}_{\mathtt{X}}}$ is $\mathrm{GHZ}_{n,m}$ but with $\mathtt{X}$ removed (as well as the $m + 1$ edges stemming from it).

That this map is a bona fide trick is the content of the next theorem [15].

Theorem 24. *The GHZ conditioning trick is QC-gap non-increasing for* $\mathrm{GHZ}_{n,m}$ *and* QGHZ_n.

Definition 11 (GHZ Marginalization Trick)**.** Let $\mathtt{Y}_i$ be a visible vertex in $\mathrm{GHZ}_{n,m}$ satisfying $\mathtt{pa}(\mathtt{Y}_i) \neq \varnothing$. The *GHZ marginalization trick* is the map

$$\mathrm{marg}_{\mathtt{Y}_i} : (\mathrm{GHZ}_{n,m}, \mathrm{QGHZ}_n) \mapsto (\mathrm{GHZ}_{n,m}^{\mathrm{marg}_{\mathtt{Y}_i}}, \mathrm{QGHZ}_n^{\mathrm{marg}_{\mathtt{Y}_i}}), \tag{76}$$

where $\mathrm{QGHZ}_n^{\mathrm{marg}_{\mathtt{Y}_i}} = \mathrm{QGHZ}_{n-1}$ with $\mathtt{Y}_i$ and $\mathtt{X}_i$ removed (as well as all edges including them) and $\mathrm{GHZ}_{n,m}^{\mathrm{marg}_{\mathtt{Y}_i}} = \mathrm{GHZ}_{n-1,m}$ with $\mathtt{Y}_i$ and $\mathtt{X}_i$ removed (as well as all edges including them).

As before, this map is a genuine trick because of the following theorem [15].

Theorem 25. *The GHZ marginalization trick is QC-gap non-increasing for* $\mathrm{GHZ}_{n,m}$ *and* QGHZ_n.

B Selected Proofs

B.1 Proofs for Section 2

Proof of Theorem 5. (See Appendix A for the preliminaries of this proof.) We induct on $n \geq 2$. If $n = 2$, then GHZ_2 is the Bell scenario, which by Bell's theorem admits a QC-gap. Now suppose GHZ_n admits a QC-gap and consider GHZ_{n+1}. Marginalizing $\mathtt{Y}_{n+1}$ generates a distribution compatible with GHZ_n, which by assumption admits a QC-gap. Since marginalization is QC-gap non-increasing, GHZ_{n+1} must also admit a QC-gap. ■

Proof of Theorem 6. (See Appendix A for the preliminaries of this proof.) If $n > m + 1$, then in $\mathrm{GHZ}_{n,m}$ there exist visible nodes $\mathtt{Y}_i$ and $\mathtt{Y}_j$ such that $\mathsf{pa}(\mathtt{Y}_i) \cap \mathsf{pa}(\mathtt{Y}_j) = \{\Lambda\}$. Marginalize every $\mathtt{Y}_k \in \{\mathtt{Y}_1, \ldots, \mathtt{Y}_n\} \backslash \{\mathtt{Y}_i, \mathtt{Y}_j\}$ and then condition each $\mathtt{X}_k \in \{\mathtt{X}_1, \ldots, \mathtt{X}_n\} \backslash \{\mathtt{X}_i, \mathtt{X}_j\}$ on some $x_k \in \Sigma$. This generates an mDAG that is isomorphic to $\mathrm{GHZ}_{2,0}$, which, by Bell's theorem, admits a QC-gap. Therefore, since both marginalization and conditioning are QC-gap non-increasing, $\mathrm{GHZ}_{n,m}$ with $n > m + 1$ must also admit a QC-gap. ■

B.2 Proofs for Section 3

Proof of Lemma 7. Each $\mathtt{Y}_j$ is the output of an elementary gate with fan-in at most k. Each of those inputs come from a gate with fan-in at most k. The argument repeats down to the longest path in the graph, that is, to the depth d which is constant. Thus, ignoring the advice, $\mathtt{Y}_j$ can depend on at most k^d input bits. ■

Proof of Lemma 11. It suffices to prove that $\mathsf{NC}^0/\mathsf{rpoly}_\epsilon \subseteq \mathsf{NC}^0$ holds for all $\epsilon \in (0, \frac{1}{2})$. To this end, fix $\epsilon \in (0, \frac{1}{2})$ and suppose $L \in \mathsf{NC}^0/\mathsf{rpoly}_\epsilon$. Then there exists a family of shallow circuits $(\mathcal{C}_{n+r})_{n \in \mathbb{N}}$ with $r = \mathrm{poly}(n)$ bits of randomized advice, depth d and fan-in k such that for all $x \in L$ the input $\langle x \rangle$ implies $\Pr_{\mathtt{Y}|\mathtt{X} \sim \mathscr{D}(\mathcal{C}_{|x|+r})}(\mathtt{Y} = 1 \mid \mathtt{X} = x) \geq 1 - \epsilon$.

Since each $\mathcal{C}_{n+r} : \{0,1\}^{n+r} \to \{0,1\}$ has bounded fan-in k and constant depth d, its output is a function of at most k^d inputs. Therefore, deciding $L \in \mathsf{NC}^0/\mathsf{rpoly}_\epsilon$ is equivalent to computing a k^d-ary function $f : \{0,1\}^{k^d} \to \{0,1\}$. There are 2^{k^d} such functions, each of which is computable by the

shallow circuit equal to its DNF. Thus, $L \in \mathsf{NC}^0$, and hence for all $\epsilon \in (0, \frac{1}{2})$ it holds that $\mathsf{NC}^0/\mathsf{rpoly}_\epsilon \subseteq \mathsf{NC}^0$. ■

Proof of Lemma 13. We prove the proposition

$$\begin{aligned}\left(\forall \epsilon_1, \epsilon_2 > 0 : \mathsf{SampNC}^0/\mathsf{rpoly}_{\epsilon_1} = \mathsf{SampNC}^0/\mathsf{rpoly}_{\epsilon_2}\right) \\ \implies \left(\forall \delta_1, \delta_2 \in (0, 1/2) : \mathsf{FNC}^0/\mathsf{rpoly}_{\delta_1} = \mathsf{FNC}^0/\mathsf{rpoly}_{\delta_2}\right), \quad (77)\end{aligned}$$

from which the result follows by contrapositive and Lemma 9.

Supposing the premise of Eq. (77), choose $\delta_1, \delta_2 \in (0, \frac{1}{2})$ such that $\delta_1 < \delta_2$. (If $\delta_1 = \delta_2$, then we are done). Evidently, $\mathsf{FNC}^0/\mathsf{rpoly}_{\delta_1} \subseteq \mathsf{FNC}^0/\mathsf{rpoly}_{\delta_2}$, so it remains to prove the reverse direction. To this end, let $R = (A_x)_{x \in \{0,1\}^*}$ be a search problem in $\mathsf{FNC}^0/\mathsf{rpoly}_{\delta_2}$. Then there exists a shallow circuit family $(\mathcal{C}_{n+r})_{n \in \mathbb{N}}$ such that for all $x \in \{0,1\}^*$ the input $\langle x \rangle$ implies

$$\mathscr{D}_{x,\delta_2} \coloneqq \Pr_{\mathtt{Y}|\mathtt{X} \sim \mathscr{D}(\mathcal{C}_{|x|+r})}(\mathtt{Y} \in A_x \mid \mathtt{X} = x) \geq 1 - \delta_2. \quad (78)$$

The collection $S = (\mathscr{D}_{x,\delta_2})_{x \in \{0,1\}^*}$ defines a sampling problem. Evidently, $S \in \mathsf{SampNC}^0/\mathsf{rpoly}_{\delta_2}$ because $\left\|\mathscr{D}(\mathcal{C}_{|x|+r}) - \mathscr{D}_{x,\delta_2}\right\| = 0$. By assumption, $\mathscr{D}_{x,\delta_2} \in \mathsf{SampNC}^0/\mathsf{rpoly}_\delta$ for any $\delta > 0$, which means there exists a shallow circuit family $(\widetilde{\mathcal{C}}_{n+\tilde{r}})_{n \in \mathbb{N}}$ such that for all $x \in \{0,1\}^*$ the input $\langle x \rangle$ implies

$$\left\|\mathscr{D}(\widetilde{\mathcal{C}}_{|x|+\tilde{r}}) - \mathscr{D}_{x,\delta_2}\right\| \leq \delta. \quad (79)$$

Thus, $(\widetilde{\mathcal{C}}_{n+r})_{n \in \mathbb{N}}$ samples from $\mathscr{D}_{x,\delta_2}$ to within any $\delta > 0$, which is to say that $(\widetilde{\mathcal{C}}_{n+r})_{n \in \mathbb{N}}$ solves R. Hence, $R \in \mathsf{FNC}^0/\mathsf{rpoly}_{\delta_1}$, as desired. ■

Proof of Lemma 15. Let $S_{\text{uniform}} \coloneqq (\mathscr{D}_x)_{x \in \{0,1\}^*}$ be the *uniform sampling problem*, where each $\mathscr{D}_x$ is a uniform distribution over $p(|x|) \in \text{poly}(|x|)$ bits—that is, $\Pr_{\mathtt{Y} \sim \mathscr{D}_x}(\mathtt{Y}) = 1/p(|x|)$. Consider now the family of classical shallow circuits $(\mathcal{C}_{n+r})_{n \in \mathbb{N}}$ with $r = p(x)$ bits of randomized advice Λ such that for all $x \in \{0,1\}^*$ it holds that $\mathcal{C}_{|x|+r}(x, \Lambda) = \Lambda$. In other words, the input is simply ignored. This circuit family has zero depth and zero fan-in, so it is trivially shallow. Now choose the randomized advice to be $\mathscr{D}_x$-random. Then the output of each $\mathcal{C}_{|x|+r}$ samples from $\mathscr{D}_x$. Hence, for all $x \in \{0,1\}^*$ the input $\langle x \rangle$ implies $\left\|\mathscr{D}(\mathcal{C}_{|x|+r}) - \mathscr{D}_x\right\| = 0$, and therefore $S_{\text{uniform}} \in \mathsf{SampNC}^0/\mathsf{rpoly}$.

To obtain a contradiction, suppose $S_{\text{uniform}} \in \mathsf{SampNC}^0$. Then for all $\epsilon > 0$ there exists a family of shallow circuits $(\mathcal{C}_n)_{n\in\mathbb{N}}$ such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies $\left\|\mathscr{D}(\mathcal{C}_{|x|}) - \mathscr{D}_x\right\| \leq \epsilon$. This implies, for every valuation $\mathtt{Y} = y \in \{0,1\}^{p(|x|)}$,

$$\frac{1}{2}\left|\Pr_{\mathtt{Y}\sim\mathscr{D}(\mathcal{C}_{|x|})}(\mathtt{Y}) - \frac{1}{p(|x|)}\right| \leq \epsilon. \tag{80}$$

But this is impossible, because $\mathcal{C}_{|x|}$ is deterministic, and so all but one valuation is zero. Therefore, $S_{\text{uniform}} \notin \mathsf{SampNC}^0$. ■

Proof of Lemma 19. Suppose $\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon = \mathsf{SampQNC}^0$. It is plain that $\mathsf{FNC}^0/\mathsf{rpoly}_\epsilon \subseteq \mathsf{FQNC}^0$, so it remains to prove $\mathsf{FQNC}^0 \subseteq \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon$. To this end, let $R = (A_x)_{x\in\{0,1\}^*}$ be a search problem in FQNC^0. Then for all $\delta > 0$ there exists a family of quantum shallow circuits $(\mathcal{Q}_n)_{n\in\mathbb{N}}$ such that for all $x \in \{0,1\}^*$ the input $\langle x\rangle$ implies

$$\mathscr{D}_{x,\delta} \coloneqq \Pr_{\mathtt{Y}|\mathtt{X}\sim\mathscr{D}(\mathcal{Q}_{|x|})}(\mathtt{Y} \in A_x \mid \mathtt{X} = x) \geq 1 - \delta. \tag{81}$$

The collection $S = (\mathscr{D}_{x,\delta})_{x\in\{0,1\}^*}$ constitutes a sampling problem. Evidently, $S \in \mathsf{SampQNC}^0$ because $\left\|\mathscr{D}(\mathcal{Q}_{|x|}) - \mathscr{D}_{x,\delta}\right\| = 0$. Thus, by the assumption $\mathsf{SampNC}^0/\mathsf{rpoly}_\epsilon = \mathsf{SampQNC}^0$, there is a family of classical shallow circuits that sample from the distributions in S to within ϵ. This is sufficient, because by construction the distributions in S solve the search problem R. ■

Proof of Lemma 20. The proof is the same as the proof of Lemma 19, except with the definitions mended for proving the statement $\mathsf{FQNC}^0_\epsilon \subseteq \mathsf{FNC}^0/\mathsf{rpoly}_\epsilon$. ■