

Complexity-Theoretic Foundations of BosonSampling with a Linear Number of Modes

Adam Bouland,¹ Daniel Brod²,³ Ishaun Datta¹, Bill Fefferman³, Daniel Grier^{4,5},
Felipe Hernández⁶, and Michał Oszmaniec⁷

¹Department of Computer Science, *Stanford University*, Stanford, California 94305, USA

²Instituto de Física, *Universidade Federal Fluminense*, 24210-346 Niterói, Rio de Janeiro, Brazil

³Department of Computer Science, *University of Chicago*, Chicago, Illinois 60637, USA

⁴Department of Computer Science and Engineering, *UC San Diego*, La Jolla, California 92093, USA

⁵Department of Mathematics, *UC San Diego*, La Jolla, California 92093, USA

⁶Department of Mathematics, *MIT*, Cambridge, Massachusetts 02139, USA

⁷Center for Theoretical Physics, *Polish Academy of Sciences*, 02-668 Warszawa, Poland



(Received 2 April 2025; revised 10 December 2025; accepted 5 March 2026; published 24 June 2026)

BosonSampling is the leading candidate for demonstrating quantum computational advantage in photonic systems. While we have recently seen many impressive experimental demonstrations, there is still a formidable distance between the complexity-theoretic hardness arguments and current experiments. One of the largest gaps involves the ratio of particles to modes—all current hardness evidence assumes a “dilute” regime in which the number of linear optical modes scales at least quadratically in the number of particles. By contrast, current experiments operate in a “saturated” regime with a linear number of modes. In this paper, we bridge this gap, bringing the hardness evidence for experiments in the saturated regime to the same level as had been previously established for the dilute regime. This involves proving a worst-to-average-case reduction for computing the permanent, which is robust both to large numbers of row repetitions and also to distributions over matrices with correlated entries. We also apply similar arguments to give evidence for hardness of Gaussian BosonSampling in the saturated regime.

DOI: [10.1103/xc7b-sjm5](https://doi.org/10.1103/xc7b-sjm5)

Subject Areas: Photonics,
Quantum Information

I. INTRODUCTION

In the decade since it was proposed by Aaronson and Arkhipov [1], BosonSampling has become one of the most promising candidates for achieving quantum computational advantage—the experimental demonstration of a quantum computation which exponentially surpasses classical computers. This requires a task which both is experimentally feasible and has strong complexity-theoretic evidence for hardness. We have now seen several experimental demonstrations of BosonSampling [2,3] and its Gaussian variant at scale [4–6], as well as substantial work building the theory of these experiments and bringing them closer to feasibility (see, e.g., Refs. [7–10]).

Despite this progress, there is still a formidable distance between experiment and theory. One of the most notable gaps involves the ratio between the number of modes and number of particles (typically, photons). The original BosonSampling proposal calls for a “dilute regime” in

which n photons are passed through an $m = \Omega(n^2)$ -mode Haar-random interferometer followed by the measurement of each mode in the particle-number basis. More precisely, the original paper required $m = \omega(n^5)$ modes but showed that $m = \Omega(n^2)$ would suffice under a plausible random matrix theory conjecture. Subsequent work proposed a variant, known as bipartite Gaussian BosonSampling [9], that works with $\Omega(n^2)$ modes. By contrast, all experiments to date operate in a “saturated regime” where the number of modes is comparable to the number of photons, i.e., $m = \Theta(n)$. The ratio m/n varies from 3 to 5.6 in recent BosonSampling experiments [2,3] and from 0.5 to 2 in recent Gaussian BosonSampling experiments [4–6,11]. Understanding this saturated regime has long been cited as a major open problem going back to the original paper [1], which asked explicitly: “*Can we reduce the number of modes needed for our linear-optics experiment, perhaps from $O(n^2)$ to $O(n)$?*”

It is reasonable to conjecture that the need for such a high number of modes is an artifact of current proof techniques rather than intrinsic to the hardness of the sampling problem. For one, state-of-the-art classical simulation algorithms are not able to take advantage of the saturated regime to achieve dramatically faster run times [12], albeit

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

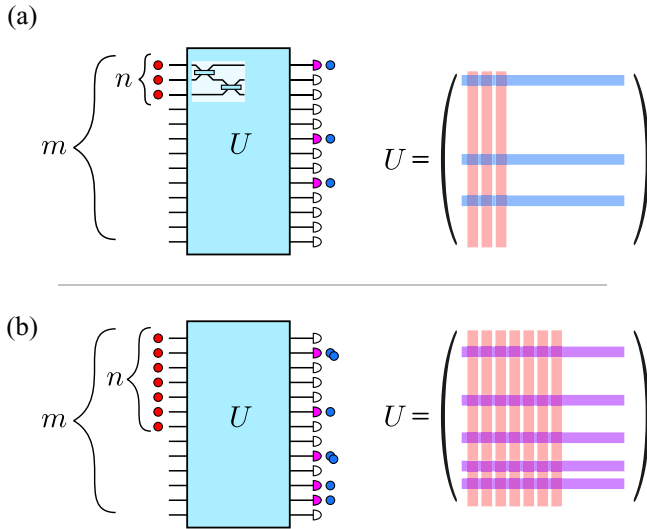


FIG. 1. Two regimes of BosonSampling. (a) The dilute limit, where $m = \Omega(n^2)$. No-collision outcomes dominate, and the submatrix whose permanent we must compute (i.e., selecting the appropriate rows and columns) is sufficiently smaller than the Haar-random interferometer U that its entries look i.i.d. Gaussian. (b) The saturated limit, where $m = \Theta(n)$. No-collision outcomes are rare, and complexity is determined by the number of detector clicks (purple) rather than output photons (blue). The entries of the submatrix of U are very correlated, and its rows will be repeated according to collisions, as described in the main text.

some improvements are possible [13,14]. Furthermore, the saturated regime is sufficient to perform universal quantum computation [15] with single-photon sources, linear optics, and (adaptive) photon detectors.

Analyzing the hardness of BosonSampling in the saturated regime is quite challenging for two reasons (see Fig. 1 for a graphical presentation). The first is due to collisions, i.e., events where more than one photon is detected in a given output mode. Current proof techniques rely heavily on a property of the dilute regime known as the “bosonic birthday paradox” [16], which ensures that most measurement outcomes are collision-free. By contrast, BosonSampling in the saturated regime has a large number of collisions, which yields many inequivalent output types that have to be independently analyzed. Second, the probability of each outcome is the squared permanent of a submatrix of the unitary that encodes the interferometer, which is chosen to be Haar random. In the dilute limit, these submatrices have entries that are approximately i.i.d. (independent and identically distributed) Gaussian, which are convenient to analyze, whereas in the saturated limit, the relevant submatrices do not have i.i.d. entries [17].

In this work, we overcome these obstacles and build the complexity-theoretic foundations for BosonSampling in the saturated regime, answering Aaronson and Arkhipov’s question in the affirmative. The starting point is to identify typical outputs that occur in generic

BosonSampling experiments in the saturated regime. Specifically, we prove that typical outcomes in n -particle and m -mode BosonSampling experiments are characterized by $c \approx [(m/n)/(m/n + 1)]n$ clicks (i.e., modes occupied by at least a single particle). Therefore, in the saturated regime, we still have $c = \Theta(n)$. Then, using methods from Ref. [9], we show that probabilities of such typical outcomes can be used to encode permanents of square matrices of size c . This connection allows us to apply the complexity-theoretic framework of Aaronson and Arkhipov to prove that hardness of classical approximate sampling from BosonSampling in the experiments follows from the hardness of an appropriate average-case hardness conjecture.

Main result (informal)—Assuming average-case hardness of approximating random output probabilities of BosonSampling experiments for $m \geq 2.1n$, there is no efficient classical algorithm to sample from the output distribution of such an experiment. Furthermore, we prove a weaker version of this average-case hardness conjecture. The gap between what is proven and what is conjectured is the same as for standard BosonSampling in the dilute regime.

Remark. While for concreteness we state our results for $m > 2.1n$ throughout this paper, in fact, all such statements pertain more generally to $m = \lceil (2 + \epsilon)n \rceil$ and $n \geq 2\epsilon^{-1}$ for any $\epsilon > 0$. Though our result does not formally hold when $n \leq m \leq 2n$, we believe this to be not fundamental but rather an artifact of the proof techniques used in Lemma B2.

Our main results give strong evidence in favor of this average-case hardness conjecture. In particular, we show it is #P-hard to approximately compute random output probabilities of random experiments in the saturated regime, albeit to a smaller error than would be necessary. Additionally, we extend the arguments based on Stockmeyer reduction to the case where hiding does not hold; i.e., not all outcomes in random experiments are equivalent. We also provide numerical evidence for anti-concentration in this regime. This brings the saturated regime of BosonSampling to essentially the same level of theoretical support as dilute-limit experiments. Importantly, we also apply our complexity-theoretic analysis to Gaussian BosonSampling, which is the choice of state-of-the-art experiments [4–6,11]. In this case, the saturated regime where our results formally hold (corresponding to $m > 2.1n$) matches to a squeezing parameter of 0.65, which is well within the range of experimentally accessible values [5].

Beyond reinforcing the theoretical underpinnings of saturated regime of BosonSampling, our work also has consequences for experimental demonstrations of BosonSampling. First, a combinatorial characterization of number of detector clicks in the saturated regime can serve as a benchmark for current and future experiments which operate in the saturated regime. From a physics perspective, our analysis highlights the role of *number*

of *detector clicks* as the measure of hardness of BosonSampling experiments beyond the diluted regime.

Our work is also relevant from the point of view of losses. Losses are one of the main limitations in scaling up BosonSampling experiments [18–22], particularly since losses grow exponentially with circuit depth, and circuit depth for Haar-random interferometers typically scales linearly with the number of modes. Using the value of 99.92% for state-of-the-art transmittance per optical component in a photonic chip recently reported in Ref. [23], we can estimate the overall losses for large BosonSampling experiments. For an experiment with $n = 200$ photons, the overall transmittance of the linear-optical circuit (i.e., ignoring losses in other components such as detectors and in the coupling into and out of the chip) in the saturated regime where our results hold, when $m \approx 2n$, are of the order of 73%. In the dilute regime, choosing $m = n^2$, the overall transmittance drops to $10^{-12}\%$. Thus, even though losses remain a major obstacle for scaling up BosonSampling in the real world, our results highlight a sweet spot where the model is both sufficiently complex and realistic.

A. Proof sketch

We begin by reviewing the well-known connection between linear-optical experiments and the permanent function [1,24]. This will serve as background for the unfamiliar reader but also to lay out more concretely how the new challenges which we attributed to the saturated regime arise.

Suppose we have m photonic modes on which we prepare a Fock state of $n < m$ photons. Each photon is initialized in a different mode, and without loss of generality we assume that the first n modes are initially occupied. These modes are then subject to a linear-optical evolution (or interferometer), i.e., a transformation of the type

$$a_j^\dagger \mapsto \sum_i U_{ij} a_i^\dagger, \quad (1)$$

where $a_j^\dagger$ is the bosonic creation operator corresponding to mode j and U is an $m \times m$ unitary matrix. Finally, we detect the number s_i of photons that exit on the i th output mode of the interferometer. The transition probability from input $|S_0\rangle = |1, 1, \dots, 1, 0, \dots, 0\rangle$ to output $|S\rangle = |s_1, s_2, \dots, s_m\rangle$ is given by

$$p_U(S) := \frac{|\text{Per}(U_S)|^2}{\prod_{i=1}^m s_i!}, \quad (2)$$

where U_S is an $n \times n$ matrix constructed by taking the leftmost n columns of U and then subsequently taking s_i copies of the i th row.

Equation (2) is the core connection between linear-optical experiments and the permanent function. It

showcases more concretely the new challenges of the saturated regime: (i) the lack of a bosonic birthday paradox implies that, for typical U , collisions occur frequently in the output probability distribution—this, in turn, implies that the relevant matrices U_S almost surely have repeated rows—and (ii) the fact that U_S is a relatively large submatrix of U , and we cannot argue that its entries look i.i.d. Gaussian when U is uniformly random. Both challenges break major proof steps of the original BosonSampling proof [1] which we need to adapt, so now we discuss each in turn.

The first obstacle—the presence of photon collisions in typical outcomes—breaks a property known as *hiding*. Hiding is the property that all outputs of the experiment are symmetrical over the choice of random experiment. This simple property plays a surprisingly key role in current quantum advantage arguments. The basic reason is that these arguments try to show no approximate classical sampler exists to small total variation distance error (over the full outcome space). If all outputs are on equal footing, then this error can be spread over all outputs by Markov’s inequality. However, if only a few outputs matter for the hardness arguments, one might worry the approximate sampler could corrupt these outputs only—producing a distribution that is close in total variation distance to the ideal one while containing no information about the actually hard outcomes. This symmetry property trivially holds for random circuit sampling [25], circuits of commuting gates (IQP) [26], fermion sampling [27], and many other advantage schemes, as well as for BosonSampling in the dilute regime. However, in the saturated regime, this symmetry fails spectacularly—the output space shatters into an exponential number of incomparable output types.

Arguments for hardness of quantum advantage schemes [1,26,27] usually employ the Stockmeyer approximate counting algorithm [28]. This method, which crucially uses hiding, is used to prove that access to an efficient classical simulator implies the ability to approximate probabilities of *specific* outputs up to exponentially small error within the third level of the polynomial hierarchy. We instead proceed by formulating a modified version of the Stockmeyer counting reduction [1,28] which does not require the hiding symmetry. We choose a uniformly random outcome of the experiment and use Stockmeyer counting to estimate the probability of this outcome. By Markov’s inequality, we can ensure that most output probabilities of the approximate sampler are approximately correct. However, this modified reduction comes at a cost—to show hardness of sampling, it no longer suffices to show hardness of computing output probability $p_U(S)$ for a single type of S , but instead we must now show an *entire suite* of hardness results for *most* outputs S of the experiments. More formally, in the saturated limit, the output space consists of an exponential number of incomparable collision patterns—i.e., unordered lists of occupation numbers of modes. We need to argue that it is hard to

estimate the output probabilities $p_U(S)$ for *typical* S chosen from the uniform measure in the set of possible outcomes.

Our next step is to show such a suite of average-case hardness results of computing most outputs of BosonSampling experiments in the saturated regime.

Theorem 1 (Informal). It is #P-hard to compute most output probabilities of most BosonSampling experiments in the saturated regime to within additive error $e^{-\Theta(n \log n)}$.

This is nearly what we need to show hardness of sampling via the modified Stockmeyer reduction. By adapting the results of Ref. [29] to our case, it can be shown that robustness to additive error $e^{-\Theta(n)}$ would remove the need for additional hardness conjectures. To prove Theorem 1 (formally stated as Theorem 6 given in Sec. V), we need to overcome both of the major differences that distinguish the dilute and saturated regimes.

First, one must deal with the presence of collisions. To do this, we identify a collection of output types which cover a large fraction of the output probability distribution of typical experiments. This requires a careful combinatorial accounting of typical collision patterns in typical outputs of experiments in the saturated limit. In particular, our argument suggests that the main figure of merit of complexity in the saturated regime is not the number of photons or modes in the experiment but rather the number of output detector clicks (i.e., output modes which receive at least one photon), which informally follows from the fact that, if we use only $O(1)$ distinct rows repeated many times, the resulting $n \times n$ matrix has very low rank and its permanent is, in fact, easy to compute. Thus, we need to show that, in a typical experimental run, the number of clicks is still large enough even if collision-free states are rare (this is covered in Lemma 2).

Once a suitable collision pattern is identified, we need to show average-case hardness for computing output probabilities of that collision type over the random choice of interferometer. This is equivalent to showing hardness of computing the permanent of a large submatrix of a random unitary, with a particular pattern of repeated rows.

In the dilute regime, this average-case hardness argument proceeds as follows (using a variant of Lipton’s argument [30]). It starts by arguing that a Gaussian matrix is perturbed only slightly by shifting and rescaling; i.e., if B is a matrix of i.i.d. Gaussian elements, then $B(t) = (1 - t)B + tY$ is also approximately Gaussian under some mild conditions of Y and for sufficiently small t . The polynomial nature of the permanent function implies then that we can use polynomial interpolation to estimate the permanent of Y from a number of estimates of the permanent of $B(t)$. Put more simply, we can, in some sense, “sneak” a tiny amount of a worst-case matrix into an average-case matrix. Since this proof uses an entry-by-entry analysis of the matrix, it breaks in the saturated case, since the relevant submatrices are far from i.i.d. Gaussian

[17]—instead, the entries come from a highly correlated measure. We show that, perhaps surprisingly, this highly correlated distribution is nonetheless also approximately shift-and-scale invariant so that we recover Lipton’s proof. To do this, we directly study the probability density of the singular values of a submatrix of a Haar-random unitary [31,32]. We reduce the desired invariance property to estimating the gradient of this probability density, which we show is equivalent to proving sharp tail bounds on the maximum singular value. Our desired bounds require that we go beyond generic concentration inequalities such as Levy’s lemma or log-Sobolev inequalities, and instead we derive them from high-dimensional geometry. Considering the ubiquity of the Haar measure over unitaries, we expect that this bound may be of independent interest.

Finally, we address the issue of anticoncentration of outcome probabilities of saturated experiments. Anticoncentration is a necessary ingredient for converting additive estimates of output probabilities to multiplicative estimates—which we conjecture to be hard. It remains open to prove anticoncentration for all variants of BosonSampling, but there has been partial progress in this direction [33] as well as numerical evidence for anticoncentration in the dilute regime [1]. However, as one reduces the number of modes, one might worry that anticoncentration might begin to fail, due to both the row repetitions in the submatrices and the correlations between submatrix entries. This could be an issue as many of the known attacks on quantum advantage schemes hold in the nonanticoncentration regime, e.g., constant-depth random circuits [34]. To alleviate this concern and to support our anticoncentration conjecture, we provide numerical evidence that anticoncentration holds in the saturated regime and, indeed, has very similar behavior to the i.i.d. Gaussian case.

In Fig. 2, we provide schematics of the organization of the paper, with the connections between the different steps of the proof and the corresponding sections. This paper is organized as follows. In Sec. II, we define the notation used throughout the paper. In Sec. III, we prove that efficient classical simulation of saturated-regime BosonSampling is not possible up to some complexity-theoretic conjectures. In particular, we need to formulate a new conjecture adapted to the computational problem that arises in this regime. In Sec. IV, we present an analysis of the combinatorics of collisions in typical BosonSampling experiments. In Sec. V, we provide evidence toward our conjecture by proving a weaker version of the desired result. In Sec. VI, we provide numerical evidence for anticoncentration of the relevant output probabilities. In Sec. VII, we discuss how our results can be adapted to the case of Gaussian BosonSampling in the saturated regime. Finally, in Sec. VIII, we lay out some open questions and final remarks.

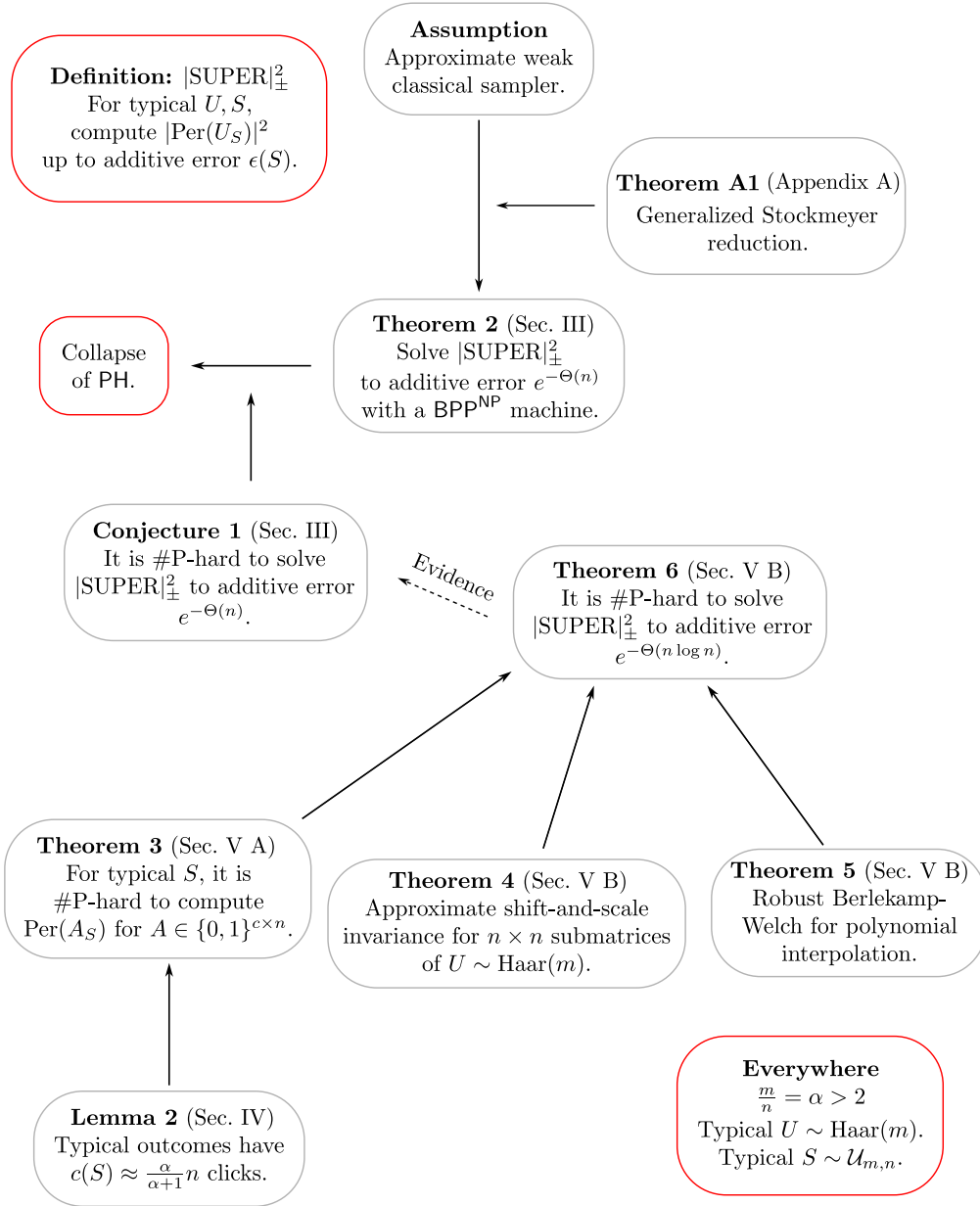


FIG. 2. Outline of the connections between new results. Formal definitions and statements of the lemmas and theorems can be found in the corresponding sections in the main text. The content of Sec. VII is organized in a similar spirit, though somewhat abbreviated, for the case of Gaussian BosonSampling.

II. NOTATION

In this section, we collect notation used throughout the paper. We use the letter n to denote the number of photons and m the number of modes. We use $\mathcal{S}_{m,n}$ for the set of possible outcome patterns $S = (s_1, \dots, s_m)$, where s_j denotes the number of photons measured in the j th mode. More precisely, we define

$$\mathcal{S}_{m,n} = \left\{ (s_1, s_2, \dots, s_m) \mid s_j \in \mathbb{N}_{\geq 0} \text{ and } \sum_{j=1}^m s_j = n \right\}. \quad (3)$$

There is a one-to-one correspondence between elements of $\mathcal{S}_{m,n}$ and Fock states of n bosons occupying m modes. Therefore, the cardinality of $\mathcal{S}_{m,n}$ equals the dimension of the Hilbert space $\text{Sym}^n(\mathbb{C}^m)$ spanned by these states. By a stars-and-bars counting argument, the cardinality of the set $\mathcal{S}_{m,n}$ is

$$|\mathcal{S}_{m,n}| = \binom{m+n-1}{n}. \quad (4)$$

In most of the paper, we assume the standard initial state, i.e., $S_0 = (1, \dots, 1, 0, \dots, 0)$ with the first n modes occupied

each by a single boson. The number of nonzero entries in a collision pattern S is referred to as the number of *clicks*, and we denote it by c_S (this is a reference to a type of detector, known as a bucket detector [35], which literally clicks when one or more photons arrive at an output mode but cannot distinguish the actual photon number). We also denote by $k_S = n - c_S$ the number of excess particles, which, for simplicity, we refer to as collisions.

We use $\mathcal{U}_{m,n}$ to refer to the uniform distribution on $\mathcal{S}_{m,n}$, i.e., the distribution such that every outcome $S \in \mathcal{S}_{m,n}$ has probability $1/|\mathcal{S}_{m,n}|$. We denote by p_U the probability distribution of $\mathcal{S}_{m,n}$ induced by BosonSampling experiments. For unitary U and outcome S , we define matrix U_S by taking the $m \times n$ left submatrix of U and repeating the i th row s_i many times. The probability distribution $p_U(S)$ on $\mathcal{S}_{m,n}$ is given by

$$p_U(S) := \frac{|\text{Per}(U_S)|^2}{\prod_{i=1}^m s_i!}. \quad (5)$$

Finally, we use $\mathcal{D}_{m,n}$ to refer to the probability distribution on $n \times n$ matrices U_S induced by drawing $U \sim \text{Haar}(m)$ and $S \sim \mathcal{U}_{m,n}$.

We also heavily rely on the standard O , Ω , and Θ notation. For two functions $f(n)$ and $g(n)$, we say that f is $O(g)$ if $f(n)$ is upper bounded by some positive multiple of $g(n)$ in the asymptotic limit $n \rightarrow \infty$. We say that f is $\Omega(g)$ if g is $O(f)$, and we say f is $\Theta(g)$ if it is both $O(g)$ and $\Omega(g)$.

III. EFFICIENT CLASSICAL SIMULATION OF BOSONSAMPLING IN THE SATURATED REGIME COLLAPSES PH

In this section, we argue that efficient classical simulation of BosonSampling in the $m = \Theta(n)$ regime would imply a collapse of the polynomial hierarchy (PH), by a reduction from sampling (the experiment's distribution over outcomes) to computing (outcome probabilities). Our proof is based on an average-case hardness conjecture for a problem we call the subunitary permanent estimation with repetitions (SUPER). In subsequent sections, we provide evidence for the hardness of this problem.

The key ideas behind the hardness proof [1] are as follows: Imagine there exists a classical sampler for a saturated BosonSampling experiment, namely, a probabilistic polynomial-time algorithm that produces an outcome S from a distribution close in l_1 distance to the experiment's. By Markov's inequality, most of the sampler's estimates to the true outcome probabilities are reasonably accurate. However, a celebrated result due to Stockmeyer, known as the approximate counting method, estimates those very outcome probabilities in BPP^{NP} (i.e., in probabilistic polynomial time with access to an NP oracle) [28]. Stockmeyer's method exploits the fact that, unlike for a

quantum computation, the classical sampler can be treated as a deterministic algorithm that takes a random input. Combining the fact that BPP^{NP} lies within the PH, our conjecture that it is $\#P$ -hard to estimate probabilities of most outcomes from most experiments in the saturated limit, and Toda's theorem (which informally places $\#P$ outside of PH), we conclude that the existence of an efficient classical sampler collapses PH—violating a bed-rock assumption of complexity theory.

Why introduce a new conjecture to argue for the classical intractability of saturated BosonSampling? As described in Sec. I, the need for a new basis to demonstrate hardness of sampling comes from the two major differences between the dilute and saturated settings, namely, collisions and correlations. In the saturated regime, outcome probabilities are not given by permanents of submatrices of Haar unitaries as in the dilute regime; instead, the matrices of interest have repeated rows due to collisions. Moreover, regardless of collisions, the relevant matrices in saturated BosonSampling manifestly do not have approximately i.i.d. Gaussian entries. Row repetitions and correlated entries disconnect standard assumptions in the BosonSampling literature, such as the permanent-of-Gaussians conjecture, from the saturated regime. To this end, we define the following problem about estimating average-case output probabilities.

Problem 1 (Subunitary permanent estimation with repetitions, $|\text{SUPER}|_{\pm}^2$). We define a family of problems parametrized by error scaling function $\varepsilon: \mathbb{N}^m \rightarrow \mathbb{R}^+$ and failure probability δ . Recall that $\mathcal{D}_{m,n}$ is the distribution on $n \times n$ matrices V obtained by drawing independently $U \sim \text{Haar}(m)$ and $S \sim \mathcal{U}_{m,n}$ and setting $V = U_S$. Given $V \sim \mathcal{D}_{m,n}$ and error probability $\delta > 0$, output $z \in \mathbb{C}$ such that $|z - |\text{Per}(V)|^2| \leq \varepsilon(S)$ with probability at least $1 - \delta$ over choice of V .

We conjecture that $|\text{SUPER}|_{\pm}^2$ is $\#P$ -hard, in particular.

Conjecture 1. $|\text{SUPER}|_{\pm}^2$ is $\#P$ -hard to additive imprecision $\varepsilon(S) = 1/[\text{poly}(n)] \prod_i s_i! / |\mathcal{S}_{m,n}|$, when $m \geq 2.1n$ with failure probability $\delta \leq \frac{1}{4}$.

Here, $\prod_i s_i!$ is a normalization factor inherited from bosonic statistics; see Eq. (5). By contrast, in dilute BosonSampling, there is only one output type and this term is trivially 1.

Theorem 2 (stated informally in Sec. I as part of the main result) shows that Conjecture 1 implies quantum computational advantage for $m = \Theta(n)$ BosonSampling.

Theorem 2 (No classical sampler). Let $m \geq 2.1n$. The existence of a classical sampler that samples from a distribution approximating the BosonSampling distribution [cf. Eq. (5)] to accuracy $\beta = 1/\text{poly}(n)$ implies a solution to $|\text{SUPER}|_{\pm}^2$ with additive imprecision $\varepsilon(S) = [1/[\text{poly}(n)]] \prod_i s_i! / |\mathcal{S}_{m,n}|$ in BPP^{NP} . Assuming Conjecture 1, this collapses PH.

We modify the usual Stockmeyer reduction to choose a uniformly random outcome. This subtlety does not arise in

dilute BosonSampling, where up to permuting the modes all outcomes are the same. The proof is deferred to Appendix A.

IV. RANDOM MEASUREMENT OUTCOMES HAVE MANY OCCUPIED MODES

For our hardness proofs, it is crucial to observe that in the saturated regime bosons do not bunch too much. That is, if $m = \Theta(n)$ for typical BosonSampling experiments, there are $\Theta(n)$ occupied modes (clicks) upon measurement. This property is important, because the bosonic transition amplitudes are proportional to permanents of matrices with repeated rows according to the collision profile. The rank of this matrix is at most the number of clicks. For a constant number of clicks, one can use efficient algorithms for constant-rank matrices, e.g., those due to Barvinok [36] and to Ivanov and Gurvits [37]. Furthermore, there exists an efficient classical algorithm to compute the permanent of an $n \times n$ matrix with only $O(\log(n))$ distinct rows (or columns) and where only a constant number of rows (columns) are repeated [18,38]. Therefore, in these high-collision regimes, transition amplitudes are easy to compute. That this is *not* the case for $m = \Theta(n)$ BosonSampling is the content of this section. Lemma 2 shows that $m/n = \alpha$ BosonSampling is dominated by outcomes with at least $[\alpha/(\alpha + 1)]n$ clicks. As we see in Sec. V, this is enough to ensure that the matrices induced by collisions in this regime inherit the hardness of computing permanents for almost all outcomes.

We are interested in the distribution of clicks c for outcomes S distributed according to the expected probability distribution of BosonSampling experiments $p_{av}(S) = \mathbb{E}_U p_U(S) = 1/|\mathcal{S}_{m,n}|$ (note that p_{av} is, in fact, the uniform distribution over $\mathcal{S}_{m,n}$). We use $\mathcal{S}_{m,n}^c$ to denote the subset of $\mathcal{S}_{m,n}$ of outcomes with exactly c clicks, i.e., c occupied modes, $c := |\{s_i | s_i \geq 1\}|$, where $c \in \{1, \dots, \min(m, n)\}$. Observe $\mathcal{S}_{m,n} = \bigcup_c \mathcal{S}_{m,n}^c$.

Lemma 1 (Combinatorics of c clicks). For $n, m \in \mathbb{N}$ and $c \in \{1, \dots, \min(m, n)\}$,

$$|\mathcal{S}_{m,n}^c| = \binom{m}{c} \binom{n-1}{n-c}.$$

Consequently, the probability that a randomly chosen $S \sim \mathcal{U}_{m,n}$ will have exactly c clicks with probability

$$p_{m,n}(c) = \frac{|\mathcal{S}_{m,n}^c|}{|\mathcal{S}_{m,n}|} = \frac{\binom{m}{c} \binom{n-1}{n-c}}{\binom{m+n-1}{n}}. \quad (6)$$

Proof. The number of states with c clicks are enumerated by first choosing which c of the m modes to “click,” i.e., allotting one boson each to c modes from among m . This contributes a factor of $\binom{m}{c}$. Next, we allot the remaining $n - c$ bosons to the c selected modes. The factor for this contribution

equals $|\mathcal{S}_{c,n-c}|$, and the dimension is given by the standard stars-and-bars argument as $\binom{(n-c)+c-1}{n-c} = \binom{n-1}{n-c}$. ■

To establish the typical behavior of c_S when $S \sim \mathcal{U}_{m,n}$ we observe that $p_{m,n}(c)$ is, in fact, the hypergeometric distribution $H(M, N, n, c)$ with parameters $M = m$, $N = m + n - 1$, $n = n$, and $c \in \{1, \dots, \min(m, n)\}$ (we use the notation from Ref. [39]). The mean of c distributed in this manner equals $\mathbb{E}[c] = (M/N)n = [\alpha/(\alpha + 1 + 1/n)]n$, where $\alpha = m/n$. Using large deviation bounds established in Ref. [39], we get strong quantitative bounds on the probability that c_S is far from the average.

Lemma 2 (Concentration inequality for number of occupied modes for random). Let c_S be a number of clicks in a random outcome $S \sim \mathcal{U}_{m,n}$ [i.e., c_S is distributed according to distribution $p_{m,n}$ in Eq. (6)]. Then, for $t \geq 0$ and $\alpha = m/n$, we have

$$\Pr_{S \sim \mathcal{U}_{m,n}} \left[\left| c_S - \frac{\alpha}{\alpha + 1 + 1/n} n \right| \geq tn \right] \leq 2 \exp(-2t^2 n). \quad (7)$$

This shows that states with $\Theta(n)$ occupied modes dominate in the $m = \Theta(n)$ regime.

V. AVERAGE-CASE HARDNESS OF SATURATED BOSONSAMPLING PROBABILITIES

In this section, we prove a weaker version of Conjecture 1, with a smaller error tolerance. We begin in Sec. VA by proving a worst-case hardness result for permanents of binary matrices with repeated rows. In Sec. VB, we leverage that result and a polynomial-interpolation argument to prove average-case hardness.

A. Worst-case hardness of permanents with repetitions

In Sec. III, we introduced SUPER. In this section, we show that, even when we relax the unitarity constraint and consider simply submatrix permanents with repetitions, the problem is still #P-hard. That is, here we demonstrate hardness of computing permanents of arbitrary (i.e., *worst-case*) binary submatrices with average-case repetitions discussed in Sec. IV. This plays an integral role in proving the hardness of computing the permanent of *average-case* submatrices with average-case repetitions, as it forms the base problem of our worst-to-average-case reduction in Sec. VB.

Theorem 3 (#P-hard permanents of worst-case submatrices). Let $m = \Theta(n)$ and $S \sim \mathcal{U}_{m,n}$. Let c_S denote the number of clicks in S . Then, with probability at least $1 - \exp[-\Theta(n)]$, there exist a poly-time-constructible matrix $A \in \{0, 1\}^{c_S \times n}$, where $c_S = \Theta(n)$ and such that it is #P-hard to compute $\text{Per}(A_S)$, where A_S is obtained from A by repeating its rows according to collision pattern S .

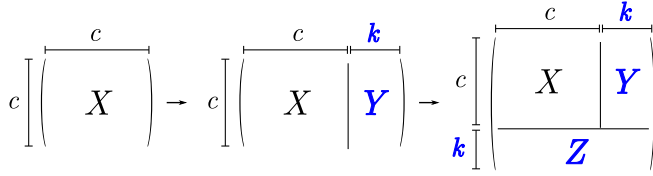


FIG. 3. The figure shows how to extend the $c \times c$ matrix X to handle k row repetitions. The first step is to create the matrix $A(X) = (X|Y)$ by appending k extra columns, as described in the text. The matrix Z just corresponds to all extra rows that are added according to the repetition pattern.

The proof of Theorem 3 uses Lemma 2, which guarantees that with high probability $c_S = \Theta(n)$ once $m = \Theta(n)$, and the following reduction between the problem of computing of permanents of binary matrices with repeated rows and the problem of computing permanents of general binary matrices.

Lemma 3. Let $m \geq n$. Given collision pattern $S \in \mathcal{S}_{m,n}$ with c clicks and matrix $X \in \{0, 1\}^{c \times c}$, there is a poly-time-constructible matrix $A(X) \in \{0, 1\}^{c \times n}$ such that

$$\text{Per}[A_S(X)] = \text{Per}(X) \prod_{i=1}^c s_i! \quad (8)$$

Lemma 3 follows directly from Lemma 24 in Ref. [9]. The general idea is as follows. We start with $c \times c$ matrix X , which encodes some #P-hard problem. We append to it a $c \times k$ matrix Y to make a larger $c \times (k+c)$ matrix, $A(X)$. Finally, we repeat k rows of the latter matrix according to a repetition pattern $(s_1, s_2, \dots, s_c)$ to make a $(k+c) \times (k+c)$ matrix $A_S(X)$. The construction is shown in Fig. 3. By an appropriate choice of Y , we can prove that $\text{Per} A_S(X)$ is proportional to $\text{Per} X$. In particular, the result of Lemma 3 is obtained from Lemma 24 in Ref. [9] by replacing B'_S and B' with $A_S(X)$ and $A(X)$, A with X , and k_s with k and by setting $z = 0$ and $y_{i,j}^{(l)} = 1$.

We illustrate the construction with a small, concrete example. Consider the following 3×3 matrix X :

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & x_{1,3} \\ x_{2,1} & x_{2,2} & x_{2,3} \\ x_{3,1} & x_{3,2} & x_{3,3} \end{pmatrix}, \quad (9)$$

and suppose $S = (4, 3, 1)$; i.e., we need to repeat the first row three additional times and the second row two additional times. This corresponds to an eight-photon BosonSampling outcome where we observed $c = 3$ non-empty output modes.

We now choose a Y matrix as follows. For each *extra* copy of a row we will need to add later, Y must contain one unit-vector column with nonzero element matching that row. More concretely, if row i will be repeated $s_i - 1$ times, then Y needs to have $s_i - 1$ columns whose i th element is 1

and all others are zero. For our particular example, we build the following matrix:

$$A(X) = \left(\begin{array}{cccc|cccc} x_{1,1} & x_{1,2} & x_{1,3} & 1 & 1 & 1 & 0 & 0 \\ x_{2,1} & x_{2,2} & x_{2,3} & 0 & 0 & 0 & 1 & 1 \\ x_{3,1} & x_{3,2} & x_{3,3} & 0 & 0 & 0 & 0 & 0 \end{array} \right). \quad (10)$$

And, after the repetition of all appropriate rows,

$$A_S(X) = \left(\begin{array}{cccc|cccc} x_{1,1} & x_{1,2} & x_{1,3} & 1 & 1 & 1 & 0 & 0 \\ x_{1,1} & x_{1,2} & x_{1,3} & 1 & 1 & 1 & 0 & 0 \\ x_{1,1} & x_{1,2} & x_{1,3} & 1 & 1 & 1 & 0 & 0 \\ x_{1,1} & x_{1,2} & x_{1,3} & 1 & 1 & 1 & 0 & 0 \\ x_{2,1} & x_{2,2} & x_{2,3} & 0 & 0 & 0 & 1 & 1 \\ x_{2,1} & x_{2,2} & x_{2,3} & 0 & 0 & 0 & 1 & 1 \\ x_{2,1} & x_{2,2} & x_{2,3} & 0 & 0 & 0 & 1 & 1 \\ x_{3,1} & x_{3,2} & x_{3,3} & 0 & 0 & 0 & 0 & 0 \end{array} \right). \quad (11)$$

It is easy to check that

$$\text{Per}[A_S(X)] = 4!3!\text{Per}(X). \quad (12)$$

In the most general case, by using this choice for Y , we get that

$$\text{Per}[A_S(X)] = \text{Per}(X) \prod_{i=1}^l s_i!, \quad (13)$$

where $k = n - c$. Note that, if X is a $\{0, 1\}$ matrix, so is A_S .

One way to see why this construction works is to consider the Laplace expansion for the permanent. Recursively apply the Laplace expansion k times, eliminating the k columns which do not include X , one at a time. It becomes clear that the choices of 0's and 1's in Y are such that the only $c \times c$ submatrices of $A_S(X)$ that contribute to the expansion are $(\prod_{i=1}^l s_i!)$ identical copies of X .

We can now prove Theorem 3.

Proof of Theorem 3. We begin by drawing a uniformly random outcome S , i.e., $S \sim U_{m,n}$. By Lemma 2, with probability $1 - \exp[-\Theta(n)]$, the outcome S has $c_S = \Theta(n)$ clicks. Consequently, computing the permanent of matrices $X \in \{0, 1\}^{c_S \times c_S}$ is #P-hard by the landmark result of Ref. [40]. By Lemma 3, taking as input our randomly drawn S and a matrix $X \in \{0, 1\}^{c_S \times c_S}$, in polynomial time one can construct our “worst-case submatrix” $A(X) \in \{0, 1\}^{c_S \times n}$ in such a way that $\text{Per}[A_S(X)] = \text{Per}(X) \prod_{i=1}^l s_i!$. Thus, $\text{Per}(A_S)$ is #P-hard as well, completing the proof. ■

B. Worst-to-average-case reduction

In this section, we prove our main result, Theorem 6, that in the saturated regime $|\text{SUPER}|_{\pm}^2$ is $\#P$ -hard to additive accuracy $\exp[-O(n \log n)]$. In Sec. VA, recall we proved the hardness of computing permanents for worst-case submatrices with average-case repetitions. Here, we present a worst-to-average-case reduction that proves $\#P$ -hardness for average-case submatrices with average-case repetitions—namely, the hardness of SUPER, Problem 1. Were the robustness, meaning the additive error tolerance, of our result to be improved to $1/|\mathcal{S}_{m,n}| = \exp[-O(n)]$, it would prove quantum computational advantage conditional *only* on the noncollapse of PH. As mentioned in Sec. I, the main challenge we need to overcome is correlations between matrix entries which breaks the original proofs in the style of Refs. [1,29].

As a reminder, the overall scheme of these reductions is an interpolation argument inspired by Lipton’s self-reducibility of the permanent, which exploits its polynomial structure to show that average-case instances are as hard as in the worst case [30]. In particular, by taking a convex combination in variable t of an average-case instance and a worst-case instance, the permanent is a univariate polynomial in t . Then, by estimating values of the polynomial for small t by the average-case algorithm, one can extrapolate to $t = 1$, the permanents of which are proved hard in Sec. VA.

The interpolation is carried out by the robust Berlekamp-Welch algorithm developed in Ref. [29], which adapts to polynomials over $\mathbb{C}$ the well-known Berlekamp-Welch algorithm [41] for polynomial interpolation over finite fields. For $m > 2n'$, $\gamma \geq 1$, robust Berlekamp-Welch demonstrates that an algorithm to solve instances of Problem 1 to additive tolerance $\exp[-(\gamma + 4)n \log n - O(n)]$ implies an efficient algorithm to compute permanents of worst-case-hard matrices under BPP^{NP} reductions. Thus, we demonstrate average-case hardness of $|\text{SUPER}|_{\pm}^2$ to additive error $\exp[-5n \log n - O(n)]$. This robustness is slightly better than that known for $m = \Omega(n^2)$ BosonSampling—with an exponent of 5 rather than 6—however, on account of the smaller size of $|\mathcal{S}_{m,n}|$ in the saturated limit, the target robustness is commensurately larger, $\exp[-O(n)]$. This is because the target additive error tolerance, or robustness, to which one seeks to prove hardness is on the order of a typical output probability, i.e., $1/|\mathcal{S}_{m,n}| = 1/\binom{m}{n} = \exp[-O(n)]$. Thus, a small gap remains between the robustness we can prove and the robustness to which we conjecture hardness. Closing the robustness gap remains open for all quantum advantage schemes.

The interpolation arguments of Refs. [1,29] proceed by an entry-by-entry analysis of i.i.d. Gaussian matrices. As described in Sec. I, for BosonSampling in the saturated limit, the matrix entries instead come from a highly correlated measure. Although in this setting this seemingly crucial i.i.d. property no longer holds, we

nonetheless prove a total variation distance bound on the distribution of average-case instances under translation and dilation that allows one to recover the interpolation argument. Theorem 4 shows that, despite correlations, the distributions are perturbed only mildly under these transformations.

Our proof of Theorem 4 develops tools that, given the ubiquity of the Haar measure over unitaries across the quantum information literature, we expect may be of independent interest. Starting with the probability density of the singular values of a submatrix of a Haar-random unitary, derived first by Ref. [31], we reduce the approximate invariance property to proving sharp tail bounds on the largest singular value. However, the bounds we need are stronger than those obtainable by black-box concentration inequalities like Levy’s lemma or log-Sobolev inequalities as used in prior work [42]. Instead, we carry out a more fine-grained analysis of the maximum singular value close to 1 by directly analyzing the geometry of high-dimensional spheres.

1. Truncations of Haar-random unitaries: Approximate invariance property

As described above, in our worst-to-average-case reduction, we smuggle a tiny amount of a worst-case matrix A into a random matrix B_t like so:

$$B_t = (1 - t)B_0 + tA \quad \text{for } t \in [0, 1], \quad (14)$$

in such a way that the permanent of B_t is (by assumption in the reduction) approximately correct with high probability for small t but at $t = 1$ is the permanent of a matrix constructed from A , which is hard to compute. This is possible in the i.i.d. Gaussian setting of dilute BosonSampling, because Gaussians are perturbed only slightly in total variation distance (TVD) by shifting and rescaling in the manner above. It is not clear that this approximate shift-and-scale invariance should hold for the highly correlated measure on truncated Haar unitaries from which B_0 is drawn. Theorem 4 shows that it does.

Theorem 4 (Approximate shift-and-scale invariance). Let μ be the probability measure on $n \times n$ complex-valued matrices induced by taking $n \times n$ submatrices of a Haar-random $m \times m$ unitary matrix, with each entry scaled by $\sqrt{m}$ to have unit variance. Fixing an arbitrary “worst-case” matrix $Y \in \{0, 1\}^{n \times n}$ and taking $B_0 \sim \mu$, we define the random variable

$$B_t = (1 - t)B_0 + tY. \quad (15)$$

We let μ_t be the probability measure for the random matrix B_t . Note that $\mu_0 = \mu$.

If $m \geq (2 + \epsilon)n$, then there exists $C_\epsilon > 0$ such that

$$\|\mu_t - \mu\|_{TV} \leq C_\epsilon nmt. \quad (16)$$

In other words, the distribution of B_t is close to that of B_0 when Y is a $\{0, 1\}$ -valued matrix and $t = o[1/(nm)]$. Note that, more generally, it suffices for matrix A to have bounded entries so that $\|Y\|_{\text{HS}} \leq \beta n$ for constant β . Furthermore, we notice that the matrix B_t in Eq. (15) is not yet a random matrix of the type that appears in $|\text{SUPER}|_{\pm}^2$, which requires some of its rows to be repeated according to S . However, suppose that for each matrix in Eq. (15) we erase the final $(n - c_S)$ rows of B_t and replace them with some copies of its first c_S rows, denoting the resulting matrices by $B_t^{(S)}$, $B_0^{(S)}$, and $Y^{(S)}$. By the data processing inequality, the TV distance bound from Theorem 4 is also an upper bound on the TV distance between distribution of

$$B_t^{(S)} = (1 - t)B_0^{(S)} + tY^{(S)} \quad (17)$$

and distribution of $\sqrt{m}U_S$, for U chosen from the Haar measure on $U(m)$. Therefore, the result of Theorem 4 directly applies to the BosonSampling matrices in the saturated regime, corresponding to c_S detector clicks and $n - c_S$ collisions.

For $m \leq 2n$, one could alternately interpolate by the Cayley path [43,44] between worst- and average-case instances on the level of unitary group $U(m)$, as is done in fermion sampling [27]. However, the final error robustness is worse than that attained using current techniques, because the degree of the resulting polynomial is $\Theta(n^2)$ [as opposed to $\Theta(n)$ using the approach from the Lipton-style proof for $m > 2n$]. We leave open whether or not it is possible to obtain similar robustness in the $m \leq 2n$ regime.

We prove Theorem 4 in Appendix B.

2. Lipton-style interpolation and robust Berlekamp-Welch

In Sec. III, we showed that Conjecture 1, namely, that $|\text{SUPER}|_{\pm}^2$ is $\#P$ -hard to additive error $\exp[-O(n)]$, implies no efficient classical sampler. Here, we provide strong evidence in favor of Conjecture 1. In particular, in Theorem 6, we prove a necessary condition for the conjecture to hold.

The idea of Lipton's proof of average-case hardness of the permanent is to take an entrywise convex combination of a matrix A_S whose permanent is hard and a matrix B_0 whose permanent we assume we can compute approximately correctly for most instances, producing matrix B_t as in Eq. (15). The squared permanent of such an object is a univariate polynomial in t of degree $2n$. For small t , points on this polynomial are close to instances of $|\text{SUPER}|_{\pm}^2$. At $t = 1$ is a (squared) permanent proved $\#P$ -hard to compute in Sec. VA. Thus, to show $|\text{SUPER}|_{\pm}^2$ is $\#P$ -hard as well, one need only interpolate from points close to $t = 0$ to $t = 1$. How close to $t = 0$ our points must be so they are approximately correct with constant probability is established by Theorem 4 as $t = O(1/n^2)$. Note that Theorem 4

is being invoked for $n \times n$ matrices A_S , B_t , and B_0 that do have repetitions according to pattern S , as they must correspond to either average instances of $|\text{SUPER}|_{\pm}^2$ or its worst-case analog (cf. discussion below Theorem 4).

Robust Berlekamp-Welch, presented below as Theorem 5, provides an algorithm for interpolation over the complex field. The algorithm runs in polynomial time with access to an NP oracle, but this is sufficient as we are already proving $\#P$ -hardness under BPP^{NP} reductions. While one could alternately use standard Lagrange interpolation, this would suffice only to prove a weaker result with hardness to a smaller amount of additive error.

Theorem 5 (Robust Berlekamp-Welch [29]). Let $p(t)$ be any degree d polynomial. There is a P^{NP} algorithm for the following task:

Input—List of data points $(t_1, y_1), \dots, (t_{100d^2}, y_{100d^2})$, where the t_i are equally spaced on the interval $[0, \Delta]$ ($\Delta < 1$) and for every $i \in [100d^2]$ we have $\Pr[|y_i - p(t_i)| \geq \epsilon] < 1/4$.

Output—Estimate $z \in \mathbb{C}$ such that $|z - p(1)| \leq \epsilon e^{d \log \Delta^{-1} + O(d)}$ with probability at least $3/4$.

The error bound on the estimate to $p(1)$ arises because the maximum blowup under extrapolation from Δ to 1 for a degree d polynomial is $(1/\Delta)^d$ (cf. Remez inequality). The subleading correction $\epsilon \exp[O(d)]$ comes from the error incurred on $[0, \Delta]$ by points that are ϵ -close to the true values, by an invocation of Paturi's bound [45].

Our main result, stated informally in Sec. I as Theorem 1, is below.

Theorem 6 ($\#P$ -hardness of SUPER in the saturated regime). Let $m \geq 2.1n$. In the regime $m = \Theta(n)$, $|\text{SUPER}|_{\pm}^2$ is $\#P$ -hard under BPP^{NP} reductions to additive error $\epsilon(S) = e^{-5n \log(n) - O(n)}$ with probability at least $1 - \delta$, with $\delta = 1/\text{poly}[(n)]$.

Proof. Suppose we have an oracle $\mathcal{O}$ to solve $|\text{SUPER}|_{\pm}^2$ with a failure probability δ (over the choices of S and U). We prove that we can, therefore, exactly compute the permanent of a specific worst-case $\{0, 1\}$ -valued matrix X (as in Lemma 3) efficiently in $\text{BPP}^{\text{NP}^{\mathcal{O}}}$. Since exactly computing the permanent of a $\{0, 1\}$ matrix is $\#P$ -hard, this will complete the theorem. Note that this is a BPP^{NP} reduction, but this is not an issue since we are proving $\#P$ -hardness.

We begin by choosing a collision pattern S at random and embedding matrix $X \in \{0, 1\}^{l \times l}$ into a matrix $A_S(X)$ with repetitions as in Lemma 3. Note that S is compatible with this embedding with high probability, due to Lemma 1, as long as $l \leq \frac{1}{2} \mathbb{E}(c_S) \approx \frac{1}{2} [m/(1 + m/n)]$. The next step is to use Theorem 4, applied to the case where all matrices in Eq. (15) have repeated rows according to S [cf. the discussion below Eq. (16) surrounding Eq. (17)]. To that end, we sample a Haar-random matrix, take a suitable submatrix of it, and repeat its rows according to S to construct $B_0^{(S)}$. We then obtain $B_t^{(S)}$ by mixing $B_0^{(S)}$ with a

small amount of $Y^{(S)}$ that we set equal to $A_S(X)$, as in Eq. (17). By Theorem 4, the ensemble defined by $B_t^{(S)}$ is close to the ensemble of (rescaled) submatrices of Haar-random matrices with repeated rows. We can, therefore, use the BPP^{NP} machine to select a Haar-random matrix U such that $B_t^{(S)}$ appears as one of its submatrices (after applying the repetition pattern S). We then use $\mathcal{O}$ to compute the approximation to absolute-squared permanent of $B_t^{(S)}$. We repeat this procedure for $100d^2 = 400n^2$ values of t in range $[0, \Delta]$, always using the same S and worst-case matrix $A_S(X)$ for each matrix $B_t^{(S)}$.

By setting $\Delta = O(1/n^2)$ and invoking the data processing inequality on Theorem 4, we satisfy the requirements of Theorem 5 that most points are ϵ -close to the true values with sufficiently high probability. Given these values as input, by Theorem 5 the robust Berlekamp-Welch algorithm outputs an approximation to $|\text{Per}[A_S(X)]|^2/m^n$ that is $\epsilon \exp[4n \log(n) + O(n)]$ -close with probability at least $3/4$. Equivalently, this corresponds to an approximation that is $\epsilon \exp[5n \log(n) + O(n)]$ -close to $|\text{Per}[A_S(X)]|^2$ for $m = O(n)$.

Now recall that $A_S(X)$ is an integer matrix; therefore, an estimate of its permanent with error less than 1 is exact. Thus, $\epsilon < \exp[-5n \log(n) - O(n)]$ additive error on the squared permanents computed by $\mathcal{O}$ suffices to compute the exact permanent of $A_S(X)$.

The only outstanding issue is that this procedure chooses a collision pattern at random, whereas $\mathcal{O}$ has to be precise only with probability $1 - \delta$ over choice of S and U . It is consistent with this definition that there are some choices of S for which $\mathcal{O}$ fails for almost every U (as long as there are some other values of S where it is unusually successful to compensate), and the conditions of Theorem 5 might not be satisfied, in which case there are no guarantees on the output of the robust Berlekamp-Welch algorithm. However, note that the specific S does not matter, since Lemmas 2 and 3 imply we can embed X into some $A_S(X)$ for almost any S . Therefore, we can repeat this entire procedure several times and take the majority vote of the results to evade these pathological choices of S . Thus, with high probability, we obtain the exact permanent of X , which is #P-hard by a well-known result due to Valiant [40] and which concludes the proof. ■

We note that our proof of Theorem 6 does not directly use Theorem 3, instead invoking Lemmas 2 and 3 for didactic purposes, though it could be rewritten to depend on Theorem 3 directly.

Notably, we achieve a robustness exponent of 5 compared to 6 for $m = \Omega(n^2)$ BosonSampling. To see that, note that all of this procedure works when $m = \alpha n$ as long as, say, $\alpha \geq 2.1$. Carrying the dependence of α in the proof, we actually obtain a robustness requirement of $\epsilon \exp[4n \log(n) + n \log(\alpha) + O(n)]$. The term $n \log(\alpha)$ is $O(n)$ in the saturated limit and $n \log(n)$ when $m = n^2$.

Note, however, that the target robustness is also larger in the saturated limit in which $1/|\mathcal{S}_{m,n}| = e^{-O(n)}$, whereas when $m = n^2$ we have $1/|\mathcal{S}_{m,n}| = e^{-O[n \log(n)]}$. Closing the robustness gap altogether remains open for all quantum computational advantage proposals.

VI. ANTICONCENTRATION IN THE PRESENCE OF CORRELATIONS

In this section, we comment on anticoncentration. Roughly speaking, anticoncentration is the property that most outcome probabilities for most BosonSampling experiments are not much smaller than the inverse of the domain size. Although in existing quantum advantage arguments anticoncentration is formally neither necessary nor sufficient for hardness, this approximate “flatness” property nevertheless plays an important role. In particular, anticoncentration is a necessary ingredient to convert additive estimates of output probabilities, which we conjecture to be hard, to multiplicative estimates. In this way, it provides a “sanity check” on hardness conjectures that involve additive error. Moreover, classical easiness results often exploit failure to anticoncentrate [34,46]. While it remains open to prove anticoncentration for any variant of BosonSampling, there has been partial progress from, e.g., Ref. [33] and strong numerical evidence in the $m = \Theta(n^2)$ regime.

In Sec. I, we discussed the two new challenges that arise when $m = \Theta(n)$ relative to the dilute regime, namely, correlations in the matrix entries and row repetitions. Though these might be expected to pose a threat to anticoncentration, below, we present numerical evidence that it holds in this regime and observe behavior similar to that in the i.i.d. Gaussian case.

For $m = \Theta(n)$, the appropriate formulation of anticoncentration is as follows.

Conjecture 2 (Anticoncentration of probabilities). There exists some polynomial such that, for all $n > 0$, $m \geq n$, $m = \Theta(n)$, and $\delta > 0$,

$$\Pr_{\substack{U \sim \text{Haar}(m) \\ S \sim \mathcal{U}_{m,n}}} \left[\frac{|\text{Per}(U_S)|^2}{\prod_{i=1}^m s_i!} < |\mathcal{S}_{m,n}|^{-1} \frac{1}{\text{poly}(m, 1/\delta)} \right] < \delta. \quad (18)$$

In Fig. 4, we show numerical evidence for Conjecture 2 in the form of plots of random BosonSampling probabilities [i.e., $|\text{Per}(U_S)|^2 / \prod_{i=1}^m s_i!$]. Recall that the conjecture posits that these probabilities are on the order of $1/|\mathcal{S}_{m,n}|$ and, in particular, do not grow smaller relative to this bound as n grows. A key takeaway from the figure is that the difference between this bound and random probabilities is quite stable as n increases (shown, in particular, on the right in Fig. 4), consistent with the conjecture. If the trend we see numerically were to continue for larger permanents—we show data up to 30 photons—then this would imply the anticoncentration conjecture.

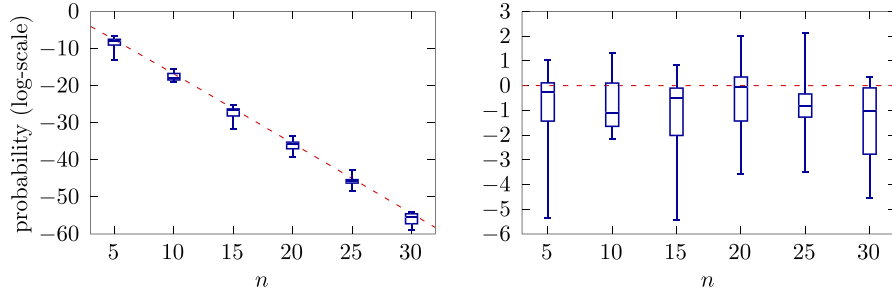


FIG. 4. Box plots for the distribution of random probabilities in the $m = 2n$ regime. For each n , we calculated $\ln[|\text{Per}(U_S)|^2 / \prod_i s_i!]$ for 20 Haar-random matrices $U \in \mathbb{C}^{2n \times 2n}$ and 20 uniformly random outcomes S . Left: the blue box plots depict this distribution, and the red dashed line shows the anticoncentration bound (i.e., $-\ln |S_{m,n}|$). Right: the blue box plots depict the same distribution shifted by the anticoncentration bound; hence, the red dashed line appears at 0. Each box plot has the same format: min, 1st quartile, median, 3rd quartile, max.

Anticoncentration motivates a reformulation of SUPER in terms of multiplicative error.

Conjecture 3. The following family of problems, denoted $|\text{SUPER}|_{\times}^2$, is #P-hard: Given $V \sim \mathcal{D}_{m,n}$ as above, output $z \in \mathbb{C}$ such that

$$|z - |\text{Per}(V)|^2| \leq [\text{poly}(m)]^{-1} |\text{Per}(V)|^2 \quad (19)$$

with probability at least $3/4$ over choice of $V \sim \mathcal{D}_{m,n}$.

Conjectures 2 and 3 immediately yield that $|\text{SUPER}|_{\pm}^2$ is #P-hard to error $\prod_i s_i! |S_{m,n}|^{-1}$, so we recover the same hardness condition as Conjecture 1, which by Sec. III again implies no classical sampler.

VII. GAUSSIAN BOSONSAMPLING

In this section, we show how our main results still apply to the BosonSampling variant known as Gaussian BosonSampling (GBS) [10]. The overall logic of the argument is the same, so we focus only on those steps of our main result that need to be modified in this case.

In Gaussian BosonSampling, rather than a Fock-state input, one uses a Gaussian state, such as a two-mode-squeezed state, as described in Refs. [8,10]. Usually, this means that the probabilities are written in terms of Hafnians of particular transition matrices rather than permanents. To sidestep having to worry about Hafnians and to reuse much of our previous work, we instead use the construction of Ref. [9], known as BipartiteGBS.

We initialize each of $2m$ modes in the same Gaussian state, $\exp[r(a^{\dagger 2} - a^2)]|0\rangle$, where r is known as the squeezing parameter. Then, this state evolves in a $2m$ -mode interferometer as shown in Fig. 5. Let us write an output state as $|S, T\rangle = |s_1, s_2, \dots, s_m, t_1, t_2, \dots, t_m\rangle$, where, due to a well-known property of two-mode squeezed vacuum states, $\sum_i s_i = \sum_i t_i = n$ (here, we consider m and n to be the number of modes and photons on either half of the interferometer rather than in total—this is convenient so that the permanents we consider are $n \times n$ as in the previous

case). The probability of observing outcome $|S, T\rangle$ is then given by

$$q(S, T) = \frac{1}{\mathcal{Z}} \frac{|\text{Per}(C_{S,T})|^2}{\prod_i s_i! \prod_j t_j!}. \quad (20)$$

In Eq. (20), $C = (\tanh r)UV^\dagger$ is a transition matrix, where U and V correspond to two m -mode interferometers and $C_{S,T}$ is a submatrix constructed from C . This submatrix follows exactly the same prescription for choosing (and repeating) rows and the standard BosonSampling one described in Sec. IA, but now we select the i th row t_i times and the j column s_j times. Finally, $\mathcal{Z} = (\cosh r)^{2m}$ is a normalization.

In our case, we choose U to be Haar random and V to be identity matrix. Under these choices, the resulting GBS model of Eq. (20) differs from the BosonSampling model we considered in Sec. IA only in a few aspects: First, the submatrices of C (which is itself proportional to a Haar-random unitary matrix) are built by repeating both rows and columns; second, the number of particles is not fixed; and,

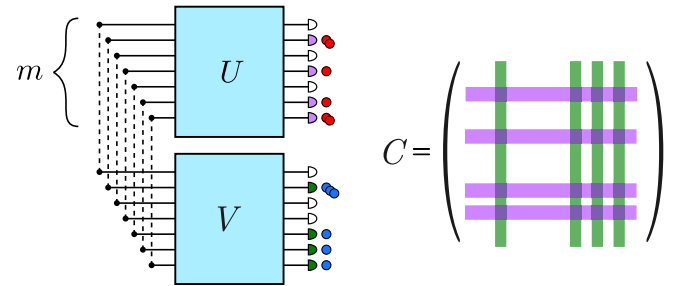


FIG. 5. Gaussian BosonSampling. Vertical dashed lines correspond to preparation of two-mode squeezed states with identical squeezing parameters r . The overall transition matrix is given by $C = (\tanh r)UW^\dagger$, and transition probabilities are given by permanents of submatrices of C , as described in the main text. Each photon observed in one of the top (bottom) m modes selects one of the rows (columns) of C .

finally, the probabilities not only are given by permanents, but are multiplied by the $1/\mathcal{Z}$ factor.

Let us first address the issue of variable photon number. Since we are assuming all squeezing parameters are the same, the number of photons emitted in each of the m top input modes independently follows a geometric with mean $\langle n_i \rangle = \sinh(r)^2$. Consequently, the total number of photons n in the top m modes follows a negative binomial distribution $P(n) = \binom{n+m-1}{n} \{[\tanh(r)^{2n}]/[\cosh(r)^{2m}]\}$ with mean $\langle n \rangle = m \sinh(r)^2$ and variance $\text{Var}(n) = m\{[\cosh(r)^2]/[\sinh(r)^4]\}$ [47]. The mode (most likely result) of the negative binomial distribution $P(n)$ equals $n_* = \lfloor (m-1) \sinh(r)^2 \rfloor$. Since r can be treated as a constant, we get that $P(n_*) \geq (C/\sqrt{m})$, where C is a constant depending only on r . This means that we can, in principle, postselect on observing exactly n_* photons on each half of the experiment by repeating the experiment only a polynomial number of times. Therefore, for simplicity, from now on we assume that we are working only in the subspace with *exactly* $n_* = \lfloor m \sinh(r)^2 \rfloor$ photons. In particular, our result holds in the regime where $m > 2.1n$, which corresponds to a squeezing parameter of roughly $r = 0.65$, well within experimentally accessible values [5].

The postselected outcome probability corresponding to m modes and n input photons in each half of the interferometer then equals

$$p_U(S, T) = \frac{1}{|S_{m,n}|} \frac{|\text{Per}(U_{S,T})|^2}{\prod_i s_i! \prod_j t_j!}. \quad (21)$$

This follows from the observation that $q(S, T) = f(n)p_U(S, T)$, where $f(n)$ is as some function of the photon number n and $p_U(S, T) = \{[|\text{Per}(U_{S,T})|^2]/(\prod_i s_i! \prod_j t_j!)\}$ can be regarded as a probability in $S \in S_{m,n}$ (for fixed $T \in S_{m,n}$) or vice versa.

We now sketch how the proof steps of previous sections would be adapted to the case of GBS.

A. Worst-case hardness of permanents with repetitions revisited

Recall that, in Sec. VA, we showed #P-hardness of worst-case permanents when rows were repeated. More concretely, we started with some $c \times c$ matrix X , we extended it into a $c \times (c+k)$ matrix $A(X)$ by appending columns according to a specific prescription, and then we repeated its rows according to some collision pattern to construct matrix $A_S(X)$. The result was Lemma 3, which showed that $\text{Per} A_S(X) = \text{Per} X \prod s_i!$. The proof of that theorem followed directly from Lemma 24 in Ref. [9] by a particular choice of all relevant variables.

Lemma 24 in Ref. [9] is, however, used to prove a stronger result, namely, Theorem 12 therein, which encompasses repetitions of both rows and columns. For our case,

this simply reduces to applying the construction of Fig. 3 twice, once for rows and once for columns.

More concretely, suppose we have two collision patterns in the output: T , for the top m modes and which selects columns, and S , for the bottom m modes and which selects rows. Let k_T and k_S be the (possibly distinct) number of *collisions* on each respective pattern. By replacing A with X , $z = 0$, and $x_{i,j}^{(l)} = y_{i,j}^{(l)} = 1$, the statement of Theorem 12 in Ref. [9] reduces to the following.

Lemma 4. Let $m \geq n$ and $S, T \in S_{m,n}$ be two collision patterns such that (i) S and T have k_S and k_T collisions, respectively, (ii) the total photon number is $n = l + k_S + k_T$, and (iii) S and T have at least k_T and k_S entries with a single photon each, respectively.

Given a matrix $X \in \{0, 1\}^{l \times l}$, there is a poly-time-constructible matrix $A(X) \in \{0, 1\}^{(l+k_T) \times (l+k_S)}$ such that

$$\text{Per}[A_{S,T}(X)] = \text{Per}(X) \prod_{i=1}^l s_i! t_i!. \quad (22)$$

Note that condition (iv) is guaranteed as long as $l \geq k_T, k_S$.

Note that the fact that we need to repeat both rows and columns adds a small overhead to the size of these matrices. In particular, whereas in the case of BosonSampling we started with a $c \times c$ matrix X , where c was the number of detectors clicks, now we construct $(l+k_T) \times (l+k_S)$ matrix whose rows and columns are then repeated, implying that we have $l+k_S$ clicks on the top side and $l+k_T$ clicks on the bottom, for a total of $n := l + k_S + k_T$ photons on each side. Furthermore, the explicit construction requires there to be k_S noncollision output modes on the T side and k_T noncollision modes on the S side. Nonetheless, from these arguments, it follows that permanents with repetitions in both rows and columns are also #P to compute in the worst case.

B. Combinatorics

We now need to prove that the analogous of Lemma 2 in the case of GBS, which would imply that *most* outcomes are hard according to Lemma 4, in the dilute regime. To that end, first note that, if we trace out either the top or the bottom half of the modes in the setup in Fig. 5, the state of the remaining modes is the maximally mixed state of n photons in m modes. This means that each half of the outcomes follows, independently, the combinatorics of Lemma 2.

One additional issue in the GBS case is that it is not enough to guarantee that there are many detector clicks on either side. As discussed just after Lemma 4, we also need the output state to have $k_S(k_T)$ modes with a single particle each on the top (bottom) half. This is likely an artifact of the proof technique of Theorem 12 in Ref. [9], but we currently do not have an alternative proof of Lemma 4 without this requirement. Nonetheless, the following lemma guarantees

that most outcome types encode a sufficiently hard problem as long as $m \geq 2.1n$.

Lemma 5. Let $m \geq 2.1n$ and let $S, T \in \mathcal{S}_{m,n}$ be collision patterns associated with the bottom and top half of our GBS setup chosen independently from $\mathcal{U}_{m,n}$. Let k_S and k_T denote the number of colliding photons in S and T , respectively. Let nc_S and nc_T denote the number of modes that are occupied and exhibit no collisions in S and T , respectively, and let c_S and c_T be the number of clicks in S and T , respectively. Then, with probability greater than $1 - \exp[-\Theta(n)]$, we have

$$k_T \leq n/3 \leq nc_S, \quad (23)$$

$$k_S \leq n/3 \leq nc_T. \quad (24)$$

Consequently, typical pairs of outputs $S, T \in \mathcal{S}_{m,n}$ can be used to compute permanents of $X \in \{0, 1\}^{l \times l}$ for $l \geq n/3$ (in the sense of Lemma 4).

Proof. We first note that we have the bound $nc_S \geq c_S - k_S$ (and similarly for output T), which follows from the fact that no collision clicks (modes) in S are the ones that are (i) occupied by a single particle and (ii) are not occupied by any of the remaining k_S “excess” particles. Using $k_S = n - c_S$ (and the same relation for outcome T), we get

$$nc_S \geq 2c_S - n, \quad nc_T \geq 2c_T - n. \quad (25)$$

Using Lemma 2 and recalling the notation $\alpha = m/n$, we get that, for any $t \geq 0$,

$$\Pr_{S \sim \mathcal{U}_{m,n}} \left[nc_S \leq \frac{\alpha - 1 + 1/n}{\alpha + 1 + 1/n} n - 2tn \right] \leq 2 \exp(-2t^2 n). \quad (26)$$

Additionally, from $k_S = n - c_S$ we get

$$\Pr_{S \sim \mathcal{U}_{m,n}} \left[k_S \geq \frac{1 + 1/n}{\alpha + 1 + 1/n} n + tn \right] \leq 2 \exp(-2t^2 n). \quad (27)$$

We have identical bounds for nc_T and k_T , where $T \sim \mathcal{U}_{m,n}$. Using the above bounds in conjunction with the union bound, it is now easy to see that, for $\alpha \geq 2.1$ and t set to a small enough constant (independent of n and m), we get that with probability at least $1 - \exp[-\Theta(n)]$ (over the choice of $S, T \sim \mathcal{U}_{m,n}$), we get inequalities (23) and (24). ■

C. Gaussian BosonSampling in the saturated limit

The big picture argument for the hardness of GBS in the saturated limit follows very closely the argument for standard BosonSampling made in the rest of the paper. We begin by positing a new problem tailored for GBS.

Problem 2 (Gaussian subunitary permanent estimation with repetitions, $|\text{GSUPER}|_{\pm}^2$). We define a family of problems parametrized by error scaling function

$\varepsilon: \mathbb{N}^m \rightarrow \mathbb{R}^+$ and failure probability δ . Let $\mathcal{G}_{m,n}$ denote a probability distribution on $n \times n$ matrices V obtained by sampling $\{S, T\} \sim \mathcal{U}_{m,n}$ and $U \sim \text{Haar}(m)$ and setting $V = U_{S,T}$. Given $V \sim \mathcal{G}_{m,n}$ and error probability $\delta > 0$, output $z \in \mathbb{C}$ such that $|z - |\text{Per}(V)|^2| \leq \varepsilon(S, T)$ with probability at least $1 - \delta$ over choice V .

The analog of Conjecture 1 for $|\text{GSUPER}|_{\pm}^2$ is the following.

Conjecture 4. $|\text{GSUPER}|_{\pm}^2$ is #P-hard to additive imprecision $\varepsilon(S, T) = \{1/\text{poly}(n)\} \prod_{i,j} s_i! t_j! / |\mathcal{S}_{m,n}|$, when $m \geq 2.1n$ with failure probability $\delta \leq \frac{1}{4}$.

We now need to prove analogs of the main two results from previously in the paper, namely, Theorems 2 and 6. Since much of the argument is similar, we focus on the differences that appear in the case of GBS.

Let us begin with the analog of Theorem 2. Recall that the gist of this theorem, which is proven in Appendix A, is that, if one has an efficient classical sampler capable of simulating a BosonSampling instance to within small total variation distance in the saturated regime, then this classical sampler can be input to Stockmeyer’s algorithm and produce a good approximation for the individual probabilities of the corresponding BosonSampling instance. This is a very generic argument, used for many quantum advantage proposals, and does not depend on any particularity of BosonSampling. The main difference that arises is that the output sample space now has dimension $|\mathcal{S}_{m,n}|^2$ rather than $|\mathcal{S}_{m,n}|$, since the GBS outputs consist of pairs of strings $\{S, T\}$ of n photons in m modes each, but this is canceled by the multiplicative factor in Eq. (21) due to how the permanent function is encoded in a GBS amplitude. Keeping track of both these changes, we obtain the following.

Theorem 7. Let $m = \Theta(n)$, $m \geq 2.1n$. An existence of a classical sampler that samples from a distribution approximating the GBS distribution [i.e., Eq. (21)] to accuracy $\beta = 1/\text{poly}(n)$ implies a solution to $|\text{SUPER}|_{\pm}^2$ with additive imprecision $\varepsilon(S, T) = \{1/\text{poly}(n)\} \prod_{i,j} s_i! t_j! / |\mathcal{S}_{m,n}|$ in BPP^{NP} . Assuming Conjecture 4, this collapses PH.

We now move to the analog of Theorem 6. Recall that, at a high level, the proof of Theorem 6 proceeds in the following steps: (i) We prove that most outcomes of a BosonSampling experiment encode a #P-hard permanent, via Theorem 3 and Lemmas 1 and 3; (ii) we prove that it is possible to smuggle a small amount of a hard matrix into a random matrix from the relevant ensemble (i.e., Theorem 4) and that then the permanent of the hard matrix is retrieved from several estimates of the permanent of random matrices via polynomial interpolation (i.e., Theorem 5); finally, (iii) the proof of Theorem 6 concatenates the previous two steps with the details of the definition of $|\text{SUPER}|_{\pm}^2$ to obtain the main result.

In the case of $|\text{GSUPER}|_{\pm}^2$, step (i) is replaced by a combination of Lemmas 4 and 5. These together show that, with high probability, the collision pattern of outcomes of a

typical GBS experiment are compatible with encoding a #P-hard problem. Step (ii) is actually unchanged, since Theorem 4 works just as well for repetitions of both rows and columns as it did for rows, and Theorem 5 just depends on the polynomial nature of the permanent. Finally, the logic of step (iii) also remains the same, though we need to keep track of numerical factors. Going through all of these in detail, we arrive at the following.

Theorem 8 (#P-hardness of GSUPER). In the regime where $m = O(n)$, $|\text{GSUPER}|_{\pm}^2$ is #P-hard to additive error $\epsilon(S) = e^{-5n \log(n) - O(n)}$, with probability at least $1 - \delta$, with $\delta = 1/\text{poly}[(n)]$.

VIII. DISCUSSION

Our work better connects the theory of BosonSampling to its experimental implementation. Aaronson and Arkhipov’s foundational work led to a number of important extensions which improved our understanding of the complexity of BosonSampling—from generalizations to Gaussian BosonSampling, to improving the robustness of average-case hardness arguments, to efficiently spoofing or verification of experiments, to characterizing the effects of noise. While many of the more empirical works have focused on the saturated regime due to its connection with experiment, most of the theoretical arguments have focused on the dilute regime and may need to be reinvestigated in this new context.

Recently, a subsequent paper of Bouland *et al.* [48] improved the robustness of hardness for dilute BosonSampling to $e^{-n \log n - n - O(n^\delta)}$ for any $\delta > 0$, i.e., within a $O(n^\delta)$ in the exponent of the desired robustness. This technique immediately applies to our results and yields robustness for Fock basis BosonSampling in the saturated regime within a factor of $O(n^\delta)$ in the exponent of the target hardness as well. However, we note other aspects of their work do not immediately carry over—such as their hardness of average-case exact sampling results—as these depend on specific properties of the Gaussian measure which do not hold in the saturated regime. We leave this as an open problem.

Many interesting questions remain. Do spoofing algorithms for BosonSampling become easier or harder in the saturated case? How few modes are needed for intractability, for example, for which α do $m = \alpha n$ experiments have the best evidence for hardness, and are there any fundamental limits on this constant?

ACKNOWLEDGMENTS

A. B. and I. D. were supported in part by the AFOSR under Grants No. FA955024-1-0089 and No. FA9550-21-1-0392. A. B. and B. F. were supported in part by the DOE QuantISED Grant No. DE-SC0020360. A. B. was supported in part by the U.S. DOE Office of Science under Award No. DE-SC0020266. D. B. acknowledges financial

support from Brazilian funding agencies FAPERJ, CNPq, and Instituto Nacional de Ciencia e Tecnologia de Informação Quântica (INCT-IQ). B. F. acknowledges support from AFOSR (FA9550-21-1-0008). M. O. was supported by the TEAM-NET project. M. O. acknowledges the support from the European Union’s Horizon Europe research and innovation program under EPIQUE Project GA No. 101135288. This material is based upon work partially supported by the National Science Foundation under Grant No. CCF-2044923 (CAREER) and by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers (Q-NEXT). F. H. was supported by the Fannie and John Hertz Fellowship. This work was done in part while A. B., I. D., B. F., D. G., and M. O. were visiting the Simons Institute for the Theory of Computing.

DATA AVAILABILITY

There are no publicly available research data or software supporting this manuscript. Requests for further information or data should be sent to the authors.

APPENDIX A: STOCKMEYER REDUCTION FOR HARDNESS OF SAMPLING

In this section, we prove Theorem 2, which shows that existence of approximate classical sampler for BosonSampling implies the solution of $|\text{SUPER}|_{\pm}^2$ problem to precision inversely proportional to $|\mathcal{S}_{m,n}|$, the dimension of symmetric subspace.

We first formalize the notion of a classical sampler.

Definition A1 (Classical sampler). A classical sampler is a probabilistic polynomial-time algorithm that, given as input $m \times m$ matrix U , and error $\beta \geq 0$, outputs a sample $S = (s_1, \dots, s_m)$ from distribution q_U such that

$$\|q_U - p_U\|_{TV} \leq \beta, \quad (\text{A1})$$

where $\mathcal{P}_U$ is the outcome distribution of the BosonSampling experiment.

As is standard among hardness of sampling results, we reduce classical sampling to approximately computing output probabilities. The reduction uses Stockmeyer’s approximate counting algorithm [28], which runs in BPP with access to an NP oracle. It is the ability in BPP^{NP} to approximate output probabilities, which are expressed in terms of squared matrix permanents we conjecture to be #P-hard, that draws a contradiction if PH is infinite.

To prove the Stockmeyer reduction, we require a technical lemma we term “reverse embedding,” which allows us to formally connect the input to $|\text{SUPER}|_{\pm}^2$, i.e., Problem 1, which is an $n \times n$ matrix drawn from the correlated distribution $\mathcal{D}_{m,n}$ of subunitaries with row repetitions, and Stockmeyer’s algorithm, which receives as input a unitary matrix via the classical sampler of

Definition A1. Reverse embedding arises whenever there is a mismatch between algorithm input types. Notably, this technicality is absent in random circuit sampling, where the input to both the classical sampler and Stockmeyer is the circuit unitary.

Lemma A1 (Reverse embedding). Given as input an $n \times n$ matrix V drawn from distribution $\mathcal{D}_{m,n}$ as in Problem 1, in BPP^{NP} one may output an $m \times m$ matrix $U = \mathcal{A}(V)$ which is distributed according to Haar measure, as well as some S such that $U_S = V$.

In particular, the distinct rows of U_S form a submatrix of the $m \times n$ truncation of U , i.e., its leftmost n columns.

Proof. In BPP, generate unitary $U \sim \text{Haar}(m)$ and $S \sim \mathcal{U}_{m,n}$, and with the NP oracle, postselect on the submatrix U_S being identical to V . By the definition of a marginal distribution, such a U and S exist. ■

Theorem A1 (Stockmeyer reduction for BosonSampling). If there exists a classical sampler as in Definition A1, then there exists in BPP^{NP} a solution to $|\text{SUPER}|_{\pm}^2$ to additive imprecision

$$\epsilon(S) = \frac{\prod_{i=1}^m s_i!}{|\mathcal{S}_{m,n}| \delta} \left(4\beta + \frac{1}{\text{poly}(m, n)} \right), \quad (\text{A2})$$

with probability at least $1 - \delta$. The numbers (s_i) come from $S = (s_1, \dots, s_m)$. Furthermore, assuming that $\beta = 1/\text{poly}(m, n)$, $\delta = \Theta(1)$, we get that there exists a BPP^{NP} solution to $|\text{SUPER}|_{\pm}^2$ with additive imprecision

$$\epsilon(S) = \frac{\prod_{i=1}^m s_i!}{|\mathcal{S}_{m,n}| \text{poly}(m, n)}, \quad (\text{A3})$$

with failure probability at most $\delta = \Theta(1)$.

Proof. Let us fix $U \in \text{U}(m)$ and observe that, because p_U and q_U are β -close in total variation distance, $p_U(S)$ and $q_U(S)$ are close with high probability for randomly chosen $S \sim \mathcal{U}_{m,n}$:

$$\mathbb{E}_{S \sim \mathcal{U}_{m,n}} |p_U(S) - q_U(S)| = \frac{1}{|\mathcal{S}_{m,n}|} \sum_{S \in \mathcal{S}_{m,n}} |p_U(S) - q_U(S)| \leq \frac{2\beta}{|\mathcal{S}_{m,n}|}. \quad (\text{A4})$$

By Markov's inequality, for any $k_1 > 0$,

$$\Pr_{S \sim \mathcal{U}_{m,n}} \left[|p_U(S) - q_U(S)| \geq \frac{2\beta k_1}{|\mathcal{S}_{m,n}|} \right] \leq \frac{1}{k_1}. \quad (\text{A5})$$

In other words, the output of the classical sampler must be accurate for most outcomes S . Now, in order to obtain a good approximation to $q_U(S)$ and, therefore, by proxy $p_U(S)$ by the above, we run Stockmeyer's algorithm that in BPP^{NP} takes as input unitary U and for any outcome S estimates $q_U(S)$ as $\tilde{q}_U(S)$ such that

$$|\tilde{q}_U(S) - q_U(S)| \leq \frac{1}{\text{poly}(m, n)} \cdot q_U(S). \quad (\text{A6})$$

The accuracy of estimate $\tilde{q}_U(S)$ depends on the magnitude of $q_U(S)$, which trivially satisfies $\mathbb{E}_{S \sim \mathcal{U}_{m,n}} [q_U(S)] \leq |\mathcal{S}_{m,n}|^{-1}$, as $q_S \leq 1$. So again by Markov, for any $k_2 > 0$,

$$\Pr_{S \sim \mathcal{U}_{m,n}} \left[q_U(S) \geq \frac{k_2}{|\mathcal{S}_{m,n}|} \right] \leq \frac{1}{k_2}. \quad (\text{A7})$$

By combining Eqs. (A5) and (A7) and employing Eq. (A6), we get

$$\Pr_{S \sim \mathcal{U}_{m,n}} \left[|p_U(S) - \tilde{q}_U(S)| \geq \frac{1}{|\mathcal{S}_{m,n}|} \left(2\beta k_1 + \frac{k_2}{\text{poly}(m, n)} \right) \right] \leq \frac{1}{k_1} + \frac{1}{k_2}. \quad (\text{A8})$$

By setting $k_1, k_2 = 2/\delta$, we get

$$\Pr_{S \sim \mathcal{U}_{m,n}} \left[|p_U(S) - \tilde{q}_U(S)| \geq \frac{1}{|\mathcal{S}_{m,n}| \delta} \left(4\beta + \frac{1}{\text{poly}(m, n)} \right) \right] \leq \delta. \quad (\text{A9})$$

Since the above inequality holds for arbitrary fixed U , we can also write

$$\Pr_{\substack{U \sim \text{Haar}(m) \\ S \sim \mathcal{U}_{m,n}}} \left[|p_U(S) - \tilde{q}_U(S)| \geq \frac{1}{|\mathcal{S}_{m,n}|\delta} \left(4\beta + \frac{1}{\text{poly}(m, n)} \right) \right] \leq \delta, \quad (\text{A10})$$

where the distributions over U and S are independent. Now, in order to connect the above estimate to the definition of $|\text{SUPER}|_{\pm}^2$, we note that (i) for randomly chosen pairs (U, S) as in Eq. (A10) we have $p_U(S) = |\text{Per}(V)|^2 / \prod_{i \in [m]} s_i!$, and, furthermore, (ii) for $V \sim \mathcal{D}_{m,n}$ we can in BPP^{NP} (cf. Lemma A1) output $U = \mathcal{A}(V)$ that is Haar distributed as well as S that is distributed according to $\mathcal{U}_{m,n}$. Consequently, it follows that

$$\Pr_{V \sim \mathcal{D}_{m,n}} \left[\left| \frac{|\text{Per}(V)|^2}{\prod_{i=1}^m s_i!} - \tilde{q}_{\mathcal{A}(V)}(S) \right| \geq \frac{1}{|\mathcal{S}_{m,n}|\delta} \left(4\beta + \frac{1}{\text{poly}(m, n)} \right) \right] \leq \delta. \quad (\text{A11})$$

We conclude the proof by noting that $(\prod_{i=1}^m s_i!) \tilde{q}_{\mathcal{A}(V)}(S)$ is the desired approximation to $|\text{Per}(V)|^2$ obtained in BPP^{NP} . ■

APPENDIX B: PROOF OF SHIFT-AND-SCALE INVARIANCE

Theorem 4 (Approximate shift-and-scale invariance). Let μ be the probability measure on $n \times n$ complex-valued matrices induced by taking $n \times n$ submatrices of a Haar-random $m \times m$ unitary matrix, with each entry scaled by $\sqrt{m}$ to have unit variance. Fixing an arbitrary “worst-case” matrix $Y \in \{0, 1\}^{n \times n}$ and taking $B_0 \sim \mu$, we define the random variable

$$B_t = (1 - t)B_0 + tY. \quad (\text{B1})$$

We let μ_t be the probability measure for the random matrix B_t . Note that $\mu_0 = \mu$.

If $m \geq (2 + \epsilon)n$, then there exists $C_\epsilon > 0$ such that

$$(\text{B2})$$

As a first step, we note that we can express the probability measure μ_θ using dilations D_ρ and translations T_H applied to the measure μ . We define these operations in terms of their actions on probability densities; we write $\mu = P(B)dB$, where $P \in L^1(\mathbb{C}^{n \times n})$ is a continuous density.

Given $\rho \in \mathbb{C}$, the dilation operator D_ρ is defined on probability densities on $\mathbb{C}^{n \times n}$ by

$$(D_\rho P)(B) = |\rho|^{-2n^2} P(B/\rho). \quad (\text{B3})$$

The factor of ρ^{-2n^2} comes from the fact that the (real) dimension of $\mathbb{C}^{n \times n}$ is $2n^2$ and ensures that $D_\rho P$ remains a probability density.

The translation T_H is defined by

$$(T_H P)(B) = P(B - H). \quad (\text{B4})$$

It is clear that $T_A P$ is the probability density of the distribution of $B + A$ when B is sampled from $P(B)dB$.

The proof of Theorem 4 is divided into two steps, given as Lemmas B1 and B2. With the definitions above, we can say that if B_0 is sampled from the probability density P , then B_t is sampled from $T_{tA} D_{1-t} P$. Lemma B1 characterizes the total variation distance between these probability distributions in terms of an integral of $|\nabla P|$, for which we prove a bound in Lemma B2.

For convenience, given matrices $A, B \in \mathbb{C}^{n \times n}$, we write $|A|$ for the Hilbert-Schmidt norm and $A \cdot B$ for the corresponding inner product. Below, it is often convenient to use $N := n^2$.

Lemma B1. Let $P \in L^1(\mathbb{C}^N)$ be any integrable function. Then, for any Y ,

$$\int |P(B) - T_Y P(B)| dB \leq \int |Y \cdot \nabla P(B)| dB. \quad (\text{B5})$$

Moreover, for any $|\delta| \leq \frac{1}{10}$

$$\begin{aligned} & \int |P(B) - D_{1+\delta} P(B)| dB \\ & \leq 10|\delta| \left(N \int |P(B)| dB + \int |B| |\nabla P(B)| dB \right). \end{aligned} \quad (\text{B6})$$

Proof. First, we prove Eq. (B3) under the assumption that P is continuously differentiable. In this case, we use the fundamental theorem of calculus in the form $f(1) - f(0) = \int_0^1 f'(t) dt$ applied to the function $f(t) = P(B - tY)$, which gives

$$P(B) - P(B - Y) = - \int_0^1 Y \cdot \nabla P(B - tY) dt. \quad (\text{B7})$$

Writing $P(B - Y) = T_A P(B)$ and using the triangle inequality above, we estimate

$$\begin{aligned}
|P(B) - T_A P(B)| &= |P(B) - P(B - Y)| \\
&\leq \int_0^1 |Y \cdot \nabla P(B - tA)| dt.
\end{aligned}$$

Integrating over B , applying Fubini's theorem, and then using the change of variables $Y = B - tY$, we have

$$\begin{aligned}
\int |P(B) - T_Y P(B)| dB &= \int \int_0^1 |Y \cdot \nabla P(B - tA)| dt dB \\
&= \int_0^1 \left(\int |Y \cdot \nabla P(B - tA)| dB \right) dt \\
&= \int_0^1 dt \int |Y \cdot \nabla P(Y)| dY. \quad (\text{B8})
\end{aligned}$$

The integral over t now evaluates to 1, and we have proved the result. The extension to all L^1 functions follows from the density of smooth functions in L^1 .

Next, we prove Eq. (B4). Fixing B , define

$$f_B(t) = |1 + t\delta|^{-2N} P[(1 + t\delta)^{-1} B].$$

Then, we compute

$$\begin{aligned}
f'_B(t) &= -2N|1 + t\delta|^{-2N-2} \text{Re}[\delta^*(1 + \delta t)] P[(1 + t\delta)^{-1} B] \\
&\quad + |1 + t\delta|^{-2N} (-\delta)(1 + t\delta)^{-2} B \cdot \nabla P[(1 + t\delta)^{-1} B].
\end{aligned}$$

In particular, using the bound $|\delta| \leq 1/10$, we obtain

$$\begin{aligned}
|f'_B(t)| &\leq 10|1 + t\delta|^{-2N} \delta \{NP[(1 + t\delta)^{-1} B] \\
&\quad + |B| |\nabla P[(1 + t\delta)^{-1} B]|\}.
\end{aligned}$$

Now, we use the triangle inequality with the fundamental theorem of calculus as before to estimate

$$\begin{aligned}
&\int |P(B) - D_{1+\delta} P(B)| dB \\
&= \int |f_B(0) - f_B(1)| dB \\
&\leq \int_0^1 \int |f'_B(t)| dB dt \\
&\leq 10\delta \int_0^1 \int |1 + t\delta|^{-2N} \{N|P[(1 + t\delta)^{-1} B]| \\
&\quad + |B| |\nabla P[(1 + t\delta)^{-1} B]|\} dB dt \\
&\leq 10\delta \int_0^1 \int N|P(B)| + |B| |\nabla P(B)| dB dt.
\end{aligned}$$

In the last line, we rescaled the variable B , which eliminated the factor of $|1 + t\delta|^{-2N}$. This concludes the proof. ■

Having expressed translation and dilation of probability density P in terms of an integral of $|\nabla P|$, we bound the integral with an explicit formula in Lemma B2.

Lemma B2. Let $P \in L^1(\mathbb{C}^{n \times n})$ be the probability density associated to the measure μ , and suppose $m \geq (2 + \epsilon)n$. Then, there exists a constant C_ϵ such that

$$\int |\nabla P(B)| dB \leq C_\epsilon \sqrt{nm} \quad (\text{B9})$$

and also

$$\int |B| |\nabla P(B)| dB \leq C_\epsilon nm.$$

Equation (B7) shows that, for $\{0, 1\}$ -valued matrices Y (for which $\|Y\|_{HS} \leq n$),

$$\int |Y \cdot \nabla P(B)| dB \leq \int \|Y\|_{HS} \|\nabla P(B)\|_{HS} dB \leq C_\epsilon nm.$$

Thus, the bounds from Lemma B2 combined with Lemma B1 (for $N = n^2$) complete the proof of Theorem 4.

The proof of Lemma B2 makes use of the following result about the largest singular values of submatrices of Haar-random unitaries, which we prove in the next subsection.

Lemma B3. If $Z_{m,n}$ is the $n \times n$ submatrix of a Haar-random $m \times m$ unitary with $m \geq (2 + \epsilon)n$, then

$$\mathbb{E}[1 - \sigma_{\max}(Z_{m,n})]^{-1} \leq C_\epsilon. \quad (\text{B10})$$

In fact, one can take $C_\epsilon = 2^{(3/\epsilon)+(7/2)}$.

Proof of Lemma B2. We first write down an explicit formula for the density P . As computed by Collins in Ref. [31] [see also Eq. (61) on p. 61 in Ref. [32] and Ref. [1]] for $m > 2n$, the density P for the measure μ is given by

$$P(B) = c_{m,n} \prod_{i \in [n]} \left(1 - \frac{\sigma_i^2}{m}\right)^{m-2n} I_{\sigma_i^2 \leq m},$$

where σ_i are the singular values of the matrix B . A first step to computing ∇P is to first take the derivatives with respect to the singular values which, for $m > 2n$,

$$\begin{aligned}
\partial_{\sigma_j} P(B) &= -c_{m,n} 2\sigma_j \frac{(m-2n)}{m} \left(1 - \frac{\sigma_j^2}{m}\right)^{m-2n-1} \\
&\quad \times I_{\sigma_j^2 \leq m} \prod_{i \neq j} \left(1 - \frac{\sigma_i^2}{m}\right)^{m-2n} I_{\sigma_i^2 \leq m} \\
&= -2\sigma_j \frac{m-2n}{m} \left(1 - \frac{\sigma_j^2}{m}\right)^{-1} P(B). \quad (\text{B11})
\end{aligned}$$

We can use this to bound $|\nabla P(B)|$. To do this, we use a singular value decomposition to reduce to the case that B is the diagonal matrix Σ with entries σ_i . The derivative of the

singular values with respect to the off-diagonal entries is 0 (it suffices to check this fact for 2×2 matrices). The same is true for the imaginary parts of the diagonal. Therefore,

$$\|\nabla P(B)\| = \left(\sum_{j \in [n]} |\partial_{\sigma_j} P(B)|^2 \right)^{1/2}. \quad (\text{B12})$$

To simplify this expression, we use $(1 - \sigma_i^2/m)^{-1} \leq (1 - \sigma_{\max}^2/m)^{-1}$ in the computation of $\partial_{\sigma_i} P(B)$:

$$\partial_{\sigma_j} P(B) \leq 2\sigma_j \left(1 - \frac{\sigma_{\max}^2}{m}\right)^{-1} P(B),$$

and, therefore, the Euclidean norm of the gradient can be bounded as follows:

$$\begin{aligned} |\nabla P(B)| &\leq 2 \left(\sum_j \sigma_j^2 \right)^{1/2} \left(1 - \frac{\sigma_{\max}^2}{m}\right)^{-1} \\ &\times P(B) \leq 2\sqrt{nm} \left(1 - \frac{\sigma_{\max}^2}{m}\right)^{-1} P(B), \end{aligned}$$

where in the last inequality we use $\sum_j \sigma_j^2 = \|B\|_{HS}^2 \leq nm$. Therefore,

$$\int |\nabla P| dB \leq 2\sqrt{nm} \left[\mathbb{E} \left(1 - \frac{\sigma_{\max}^2}{m}\right)^{-1} \right].$$

Noting that $\sigma_{\max}/\sqrt{m}$ has the law of the max singular value of Z (the unscaled submatrix of a Haar-random unitary), Lemma B3 implies that the expectation above is bounded by some C_e . Therefore, for $C'_e = 2C_e$,

$$\int |\nabla P| dB \leq C'_e \sqrt{nm}. \quad (\text{B13})$$

To estimate $\int |B| |\nabla P|$, we use that $\|B\|_{HS}^2 = \sum \sigma_j^2 \leq nm$ and, therefore,

$$|B| |\nabla P| \leq 2nm \left(1 - \frac{\sigma_{\max}^2}{m}\right)^{-1} P(B),$$

which after integrating gives

$$\int |B| |\nabla P| dB \leq 2nm \left[\mathbb{E} \left(1 - \frac{\sigma_{\max}^2}{m}\right)^{-1} \right] \leq C'_e nm.$$

This completes the proof of Lemma B2. $\blacksquare$

1. Bound on the maximum singular value of a submatrix

In this subsection, we prove Lemma B3 on the expectation $\mathbb{E}[1 - \sigma(Z_{m,n})]^{-1}$ of a submatrix of a Haar-random unitary.

In fact, we focus instead on the following bound for the probability that the maximum singular value is close to 1.

Lemma B4. Let $Z_{m,n}$ be the $n \times n$ upper left square submatrix of a Haar-random $m \times m$ unitary matrix, and let $\delta < 1/2$. Then,

$$\Pr[\sigma_{\max}(Z_{m,n}) \geq 1 - \delta] < 2^{5m-4n} \delta^{2m-4n+1}. \quad (\text{B14})$$

Before proving Lemma B4, we show how it implies Lemma B3.

Proof of Lemma B3 using Lemma B4. Using the layer-cake formula,

$$\mathbb{E}[[1 - \sigma_{\max}(Z)]^{-1}] = \int_0^\infty \Pr\{[(1 - \sigma_{\max}(Z))^{-1}] \geq t\} dt. \quad (\text{B15})$$

For $0 \leq t \leq K$, we use the fact that the integrand is bounded by 1. Then, upon rearranging and then applying Lemma B4, we have

$$\begin{aligned} \mathbb{E}[[1 - \sigma_{\max}(Z)]^{-1}] &\leq K + \int_K^\infty \Pr\{[1 - \sigma_{\max}(Z)]^{-1} \geq t\} dt \\ &= K + \int_K^\infty \Pr[\sigma_{\max}(Z) \geq 1 - t^{-1}] dt \\ &\leq K + 2^{5m-4n} \int_K^\infty t^{4n-2m-1} dt \\ &\leq K + 2^{5m-4n} \frac{1}{2m-4n} \frac{1}{K^{2m-4n}} =: F_{n,m}(K). \end{aligned} \quad (\text{B16})$$

The function $F_{n,m}(K)$ is minimized for $K_* = 2^{\frac{5m-4n}{2m-4n+1}}$ (assuming $m > 2n$). It is, moreover, easy to see that $F_{n,m}(K_*) \leq 2 \cdot 2^{(5m-4n)/(2m-4n+1)}$. If we, furthermore, use the assumption $m \geq (2 + \epsilon)n$, we can easily conclude

$$F_{n,m}(K_*) \leq 2^{(3/\epsilon)+(7/2)}. \quad (\text{B17})$$

$\blacksquare$

2. Interlude: Facts about spheres

In this section, we use ω_n to denote the volume of a ball of radius r in n dimensions. There is an explicit formula for ω_n in terms of the Gamma function:

$$\omega_n = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} r^n. \quad (\text{B18})$$

The surface area of an $(n-1)$ -dimensional sphere in n dimensions is given by $n\omega_n r^{n-1}$.

We give an elementary calculation regarding the surface area of the intersection of a sphere and a ball.

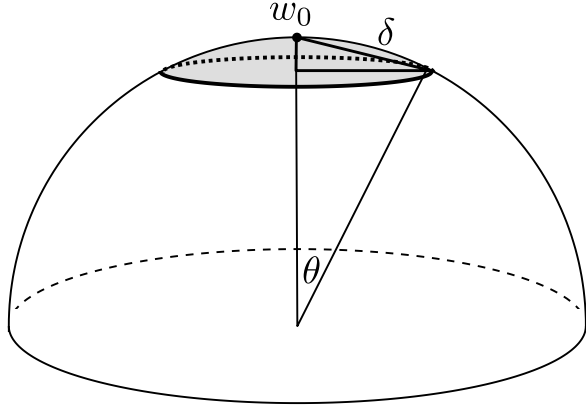


FIG. 6. The spherical cap in a δ radius around w_0 . The shaded region is a spherical cap, whose boundary is a circle (in general, $d - 2$ -dimensional sphere). The area of the cap is strictly larger than the area of the flat disk of radius $\sin \theta$ with the same boundary. The value of θ is determined by the equation $[1 - \cos(\theta)]^2 + (\sin \theta)^2 = \delta$, which can be seen by inspecting the right triangle drawn with a vertex at w_0 and a hypotenuse of length δ .

Lemma B5. Let $w_0 \in \mathbf{S}^{2n-1}$ be any unit vector and let $\delta < 1$. Then,

$$\text{Vol}_{2n-1}\{w \in \mathbf{S}^{2n-1} \mid \|w - w_0\| \leq \delta\} \geq \omega_{2n-1} 2^{-n-1/2} \delta^{2n-1}.$$

Proof. The set $\{w \in \mathbf{S}^{2n-1} \mid \|w - w_0\| \leq \delta\}$ forms a spherical cap whose boundary is a $2n - 2$ -dimensional sphere of radius $\sin(\theta)$, where θ is the angle satisfying

$$\delta^2 = (1 - \cos \theta)^2 + \sin^2 \theta.$$

See Fig. 6 for a diagram of the spherical cap. In terms of $r = \sin \theta$, which is the radius of the boundary of the spherical cap, we can write

$$\delta^2 = (1 - \sqrt{1 - r^2})^2 + r^2.$$

When $r < 1$ (which is guaranteed by $\delta < 1$), $(1 - \sqrt{1 - r^2}) < r$, so $\delta^2 \leq 2r^2$. Therefore, $r \geq \delta/\sqrt{2}$.

The volume of the spherical cap is larger than the volume of the flat disk of radius r with the same boundary, which concludes the proof. ■

Lemma B6. If $X = (x_1, x_2, \dots, x_n)$ is sampled uniformly from the unit sphere, then

$$\Pr\left(\sum_{j=1}^k x_j^2 \leq \delta\right) \leq 2^{n/2} \delta^k.$$

Proof. The first k coordinates of X are sampled from a probability distribution with density

$$p_{k,n}(x_1, \dots, x_k) = \frac{\Gamma(n/2)}{\Gamma[(n-k)/2]\pi^{k/2}} \left(1 - \sum_{j=1}^k x_j^2\right)^{(n-k)/2-1}.$$

In particular, the density is bounded by $\Gamma(n/2)/\Gamma[(n-k)/2]\pi^{k/2}$, and the event that $\sum_{j=1}^k x_j^2 \leq \delta$ is a ball of volume $\omega_k \delta^k$. Using Eq. (B16) to express ω_k , we arrive at the bound

$$\Pr\left(\sum_{j=1}^k x_j^2 \leq \delta\right) \leq \frac{\Gamma(n/2)}{\Gamma[(n-k)/2]\Gamma(k/2+1)} \delta^k.$$

Note that the prefactor involving the ratio of Gamma functions is the binomial coefficient $\binom{n/2-1}{k/2}$ and is, therefore, bounded by $2^{n/2}$, which concludes the proof. ■

3. Proof of Lemma B4

We think of the probability $\Pr[\sigma_{\max}(Z) \geq 1 - \delta]$ as the expectation of the indicator function $\mathbf{1}_{\sigma_{\max}(Z) \geq 1 - \delta}$. Though the indicator function is difficult to work with, fortunately we can bound it from above by a simpler function $f(Z)$ given by

$$\begin{aligned} f_\delta(Z) &:= \text{Vol}_{2n-1}\{w \in \mathbf{S}^{2n-1} \mid \|Zw\| \geq 1 - 2\delta\} \\ &= \int_{\mathbf{S}^{2n-1}} \mathbf{1}_{\|Zw\| \geq 1 - 2\delta} dw. \end{aligned}$$

Here, we think of an n -dimensional complex vector w as a $2n$ -dimensional real-valued vector.

We break the proof of Lemma B4 into two parts. In the first part, we show that $f_\delta(Z)$ controls the indicator function $\mathbf{1}_{\sigma_{\max}(Z) \geq 1 - \delta}$. In the second part, we compute the expectation of $f_\delta(Z)$.

For the first part, our aim is to prove that, when $\delta < 1$,

$$f_\delta(Z) \geq \omega_{2n-1} 2^{-n-1/2} \delta^{2n-1} \mathbf{1}_{\sigma_{\max}(Z) \geq 1 - \delta}. \quad (\text{B19})$$

This is equivalent to showing that, when $\sigma_{\max}(Z) \geq 1 - \delta$,

$$\text{Vol}_{\mathbf{S}^{2n-1}}\{w \in \mathbf{S}^{2n-1} \mid \|Zw\| \geq 1 - 2\delta\} \geq \omega_{2n-1} 2^{-n-1/2} \delta^{2n-1}.$$

Now suppose that $\sigma_{\max}(Z) \geq 1 - \delta$, so that, in particular, there exists a unit vector w_0 such that $\|Zw_0\| \geq 1 - \delta$. Then, for unit vectors w such that $\|w - w_0\| \leq \delta$,

$$\|Zw\| \geq \|Zw_0\| - \|Z(w - w_0)\| \geq 1 - 2\delta.$$

Therefore,

$$\begin{aligned} \text{Vol}_{\mathbf{S}^{2n-1}}\{w \in \mathbf{S}^{2n-1} \mid \|Zw\| \geq 1 - 2\delta\} \\ \geq \text{Vol}_{\mathbf{S}^{2n-1}}\{w \in \mathbf{S}^{2n-1} \mid \|w - w_0\| \leq \delta\}. \end{aligned}$$

Now Eq. (B17) follows from Lemma B5.

For the second part of the proof, we bound $\mathbb{E}f_\delta(Z)$. Using rotational symmetry we have

$$\begin{aligned}\mathbb{E}f_\delta(Z) &= \int_{\mathbb{S}^{2n-1}} \Pr(\|Zw\| \geq 1 - 2\delta) dw \\ &= \omega_{2n-1} \Pr(\|Ze\| \geq 1 - 2\delta),\end{aligned}\quad (\text{B20})$$

where $e = (1, 0, \dots, 0)$. The random vector Ze consists of the first n coordinates of the first column of a Haar-random unitary, which is simply a uniformly random vector in the unit sphere. Therefore,

$$\Pr(\|Ze\| \geq 1 - 2\delta) = \Pr\left(\sum_{j=1}^{2n} x_j^2 \geq (1 - 2\delta)^2\right)$$

when $(x_1, \dots, x_{2n})$ is sampled from the unit sphere. Rewriting the latter probability as the event that the final $2m - 2n$ coordinates are small and then applying Lemma B6, we find

$$\begin{aligned}\Pr\left(\sum_{j=1}^{2n} x_j^2 \geq (1 - 2\delta)^2\right) &= \Pr\left(\sum_{j=2n+1}^{2m} x_j^2 \leq 4(\delta - \delta^2)\right) \\ &\leq 32^{m-n} \delta^{2m-2n}.\end{aligned}$$

Combining our two bounds on $f_\delta(Z)$, we have

$$\begin{aligned}\omega_{2n-1} 2^{-n-1/2} \delta^{2n-1} \Pr[\sigma_{\max}(Z) \geq 1 - \delta] &\leq \mathbb{E}f_\delta(Z) \\ &\leq \omega_{2n-1} 32^{m-n} \delta^{2m-2n}.\end{aligned}\quad (\text{B21})$$

Lemma B4 now follows from rearranging and using $\delta < 1/2$.

[1] S. Aaronson and A. Arkhipov, *The computational complexity of linear optics*, *Theory Comput.* **9**, 143 (2013).
[2] Aaron W. Young, Shawn Geller, William J. Eckner, Nathan Schine, Scott Glancy, Emanuel Knill, and Adam M. Kaufman, *An atomic boson sampler*, *Nature (London)* **629**, 311 (2024).
[3] Hui Wang, Jian Qin, Xing Ding, Ming-Cheng Chen, Si Chen, Xiang You, Yu-Ming He, Xiao Jiang, L. You, Z. Wang *et al.*, *Boson sampling with 20 input photons and a 60-mode interferometer in a 10^{14} -dimensional Hilbert space*, *Phys. Rev. Lett.* **123**, 250503 (2019).
[4] Lars S. Madsen, Fabian Laudenbach, Mohsen Falamarzi, Askarani, Fabien Rortais, Trevor Vincent, Jacob F.F. Bulmer, Filippo M. Miatto, Leonhard Neuhaus, Lukas G. Helt, Matthew J. Collins, Adriana E. Lita, Thomas Gerrits, Sae Woo Nam, Varun D. Vaidya, Matteo Menotti, Ish Dhand, Zachary Vernon, Nicolás Quesada, and Jonathan Lavoie, *Quantum computational advantage with a programmable photonic processor*, *Nature (London)* **606**, 75 (2022).

[5] Han-Sen Zhong, Yu-Hao Deng, Jian Qin, Hui Wang, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Dian Wu, Si-Qiu Gong, Hao Su *et al.*, *Phase-programmable Gaussian boson sampling using stimulated squeezed light*, *Phys. Rev. Lett.* **127**, 180502 (2021).
[6] Han-Sen Zhong, Hui Wang, Yu-Hao Deng, Ming-Cheng Chen, Li-Chao Peng, Yi-Han Luo, Jian Qin, Dian Wu, Xing Ding, Yi Hu *et al.*, *Quantum computational advantage using photons*, *Science* **370**, 1460 (2020).
[7] L. Chakhmakhchyan and N.J. Cerf, *Boson sampling with Gaussian measurements*, *Phys. Rev. A* **96**, 032326 (2017).
[8] Abhinav Deshpande, Arthur Mehta, Trevor Vincent, Nicolás Quesada, Marcel Hinsche, Marios Ioannou, Lars Madsen, Jonathan Lavoie, Haoyu Qi, Jens Eisert, Dominik Hangleiter, Bill Fefferman, and Ish Dhand, *Quantum computational advantage via high-dimensional Gaussian boson sampling*, *Sci. Adv.* **8**, eabi7894 (2022).
[9] Daniel Grier, Daniel J. Brod, Juan Miguel Arrazola, Marcos Benicio de Andrade Alonso, and Nicolás Quesada, *The complexity of bipartite Gaussian boson sampling*, *Quantum* **6**, 863 (2022).
[10] C. S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn, and I. Jex, *Gaussian boson sampling*, *Phys. Rev. Lett.* **119**, 170501 (2017).
[11] Yu-Hao Deng *et al.*, *Gaussian boson sampling with pseudo-photon-number-resolving detectors and quantum computational advantage*, *Phys. Rev. Lett.* **131**, 150601 (2023).
[12] P. Clifford and R. Clifford, *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2018), p. 146, [10.1137/1.9781611975031.10](https://doi.org/10.1137/1.9781611975031.10).
[13] P. Clifford and R. Clifford, *Faster classical boson sampling*, *Phys. Scr.* **99**, 065121 (2020).
[14] Gregory Morse, Tomasz Rybotycki, Ágoston Kaposi, Zoltán Kolarovszki, Uros Stojic, Tamás Kozsik, Oskar Mencer, Michał Oszmaniec, Zoltán Zimborás, and Péter Rakytá, *High performance boson sampling simulation via data-flow engines*, *New J. Phys.* **26**, 033033 (2024).
[15] Emanuel Knill, Raymond Laflamme, and Gerald J. Milburn, *A scheme for efficient quantum computation with linear optics*, *Nature (London)* **409**, 46 (2001).
[16] Alex Arkhipov and Greg Kuperberg, *The bosonic birthday paradox*, *Geom. Topol. Monogr.* **18**, 1 (2012).
[17] Tiefeng Jiang, *How many entries of a typical orthogonal matrix can be approximated by independent normals?*, *Ann. Probab.* **34**, 1497 (2006).
[18] Daniel Jost Brod and Michał Oszmaniec, *Classical simulation of linear optics subject to nonuniform losses*, *Quantum* **4**, 267 (2020).
[19] Raúl García-Patrón, Jelmer J. Renema, and Valery Shchesnovich, *Simulating boson sampling in lossy architectures*, *Quantum* **3**, 169 (2019).
[20] Michał Oszmaniec and Daniel J. Brod, *Classical simulation of photonic linear optics with lost particles*, *New J. Phys.* **20**, 092002 (2018).
[21] Changhun Oh, Minzhao Liu, Yuri Alexeev, Bill Fefferman, and Liang Jiang, *Classical algorithm for simulating experimental Gaussian boson sampling*, *Nat. Phys.* **20**, 1461 (2024).

- [22] H. Qi, D.J. Brod, N. Quesada, and R. García-Patrón, *Regimes of classical simulability for noisy Gaussian boson sampling*, *Phys. Rev. Lett.* **124**, 100502 (2020).
- [23] P.I. Sund, R. Uppu, S. Paesani, and P. Lodahl, *Hardware requirements for realizing a quantum advantage with deterministic single-photon sources*, *Phys. Rev. A* **109**, 042613 (2024).
- [24] S. Scheel, *Permanents in linear optical networks*, *Acta Phys. Slovaca* **58**, 675 (2008); [arXiv:quant-ph/0406127](#).
- [25] Adam Bouland, Bill Fefferman, Chinmay Nirkhe, and Umesh Vazirani, *On the complexity and verification of quantum random circuit sampling*, *Nat. Phys.* **15**, 159 (2019).
- [26] Michael J. Bremner, Ashley Montanaro, and Dan J. Shepherd, *Average-case complexity versus approximate simulation of commuting quantum computations*, *Phys. Rev. Lett.* **117**, 080501 (2016).
- [27] Michał Oszmaniec, Ninnat Dangniam, Mauro E.S. Morales, and Zoltán Zimborás, *Fermion sampling: A robust quantum computational advantage scheme using fermionic linear optics and magic input states*, *PRX Quantum* **3**, 020328 (2022).
- [28] Larry Stockmeyer, *The complexity of approximate counting*, in *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*, STOC '83 (Association for Computing Machinery, New York, 1983), pp. 118–126.
- [29] Adam Bouland, Bill Fefferman, Zeph Landau, and Yunchao Liu, *Noise and the frontier of quantum supremacy*, in *Proceedings of the 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (2022), [arXiv:2102.01738](#).
- [30] Richard J. Lipton, *New directions in testing*, in *Distributed Computing and Cryptography*, edited by Joan Feigenbaum and Michael J. Merritt (American Mathematical Society, Oxford, 1991), Vol. 2, pp. 191–202.
- [31] Benoît Collins, *Intégrales matricielles et probabilités non-commutatives*, Ph.D. thesis, Université Pierre et Marie Curie-Paris VI, 2003.
- [32] Júlia Réffy, *Asymptotics of random unitaries*, Ph.D. thesis, BUTE Institute of Mathematics, 2005.
- [33] Sepehr Nezami, *Permanent of random matrices from representation theory: Moments, numerics, concentration, and comments on hardness of boson-sampling*, [arXiv:2104.06423](#).
- [34] J.C. Napp, R.L. La Placa, A.M. Dalzell, F.G.S.L. Brandao, and A.W. Harrow, *Efficient classical simulation of random shallow 2D quantum circuits*, *Phys. Rev. X* **12**, 021021 (2022).
- [35] Daniel J. Brod, Ernesto F. Galvão, Andrea Crespi, Roberto Osellame, Nicolò Spagnolo, and Fabio Sciarrino, *Photonic implementation of boson sampling: A review*, *Adv. Opt. Photonics* **1**, 034001 (2019).
- [36] Alexander I. Barvinok, *Two algorithmic results for the traveling salesman problem*, *Math. Oper. Res.* **21**, 65 (1996).
- [37] D. A. Ivanov and L. Gurvits, *Complexity of full counting statistics of free quantum particles in product states*, *Phys. Rev. A* **101**, 012303 (2020).
- [38] Seungbeom Chin and Joonsuk Huh, *Generalized concurrence in boson sampling*, *Sci. Rep.* **8**, 6101 (2018).
- [39] V. Chvátal, *The tail of the hypergeometric distribution*, *Discrete Math.* **25**, 285 (1979).
- [40] L. G. Valiant, *The complexity of computing the permanent*, *Theor. Comput. Sci.* **8**, 189 (1979).
- [41] Elwyn R. Berlekamp and Lloyd R. Welch, *Error correction for algebraic block codes*, US Patent No. 4,633,470, 1986.
- [42] Tanmay Singal, Filip B. Maciejewski, and Michał Oszmaniec, *Implementation of quantum measurements using classical resources and only a single ancillary qubit*, *npj Quantum Inf.* **8**, 82 (2022).
- [43] Ramis Movassagh, *The hardness of random quantum circuits*, *Nat. Phys.* **19**, 1719 (2023).
- [44] Ramis Movassagh, *Quantum supremacy and random circuits*, [arXiv:1909.06210](#).
- [45] Ramamohan Paturi, *On the degree of polynomials that approximate symmetric Boolean functions*, in *Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing*, STOC '92 (Association for Computing Machinery, New York, USA, 1992), pp. 468–474.
- [46] Boaz Barak, Chi-Ning Chou, and Xun Gao, *Spoofing linear cross-entropy benchmarking in shallow quantum circuits*, [arXiv:2005.02421](#).
- [47] R. Kruse, C. S. Hamilton, L. Sansoni, S. Barkhofen, C. Silberhorn, and I. Jex, *Detailed study of Gaussian boson sampling*, *Phys. Rev. A* **100**, 032326 (2019).
- [48] A. Bouland, I. Datta, B. Fefferman, and F. Hernández, *Exponential improvements to the average-case hardness of BosonSampling*, 2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS) (2024), pp. 912–933, [10.1109/FOCS63196.2025.00047](#).