Mitigating photon loss in linear optical quantum circuits

James Mills[1,2] and Rawad Mezher[1]

[1]Quandela, 7 Rue Léonard de Vinci, 91300 Massy, France
[2]School of Informatics, University of Edinburgh, United Kingdom

arXiv:2405.02278v4 [quant-ph] 11 Mar 2026

Photon loss rates set an effective upper limit on the size of computations that can be run on current linear optical quantum devices. We present a family of techniques designed to mitigate the effects of photon loss on both output probabilities and expectation values derived from noisy linear optical circuits composed of an input of n photons, an m–mode interferometer, and m single photon detectors. Central to these techniques is the construction *recycled probabilities*. Recycled probabilities are constructed from output statistics affected by loss, and are designed to amplify the signal of the ideal (lossless) probabilities. Classical postprocessing techniques then take recycled probabilities as input and output a set of loss-mitigated probabilities, or expectation values. Our postprocessing methods result in biased estimators of the lossless probabilities. Nevertheless, we provide both analytical and numerical evidence that these methods can be applied, up to large sample sizes, to produce output probabilities with lower combined bias and statistical errors than the statistical errors of the output probabilities obtained from postselection. Therefore, these methods can outperform postselection - currently the standard method of coping with photon loss when sampling from discrete variable linear optical quantum circuits. In contrast, we provide evidence that the popular zero-noise extrapolation technique cannot improve on the performance of postselection for any photon loss rate.

1 Introduction

Discrete variable linear optical quantum computing (DVLOQC) is a framework that uses a discrete number n of photons as well as m-mode linear optical interferometers to store and process quantum information. Many models of universal and fault-tolerant quantum computation tailored to this framework have been developed over the years beginning with the work of [1], and followed by various other models and variants (e.g [2, 3, 4, 5]). Furthermore, promising proposals for the near-term demonstration of quantum-over-classical advantage such as boson sampling [6] can naturally be implemented in this framework. A quantum device capable of performing DVLOQC usually consists of three components: single photon sources [7], multimode interferometers [8], and single photon detectors [9]. We will refer to the collection of these as a *linear optical circuit*. One major impediment to scaling up such a device is photon loss [10]. Postselection, where all *lossy* output statistics with one or more lost photons are discarded and only those where all photons are detected are kept, may be used to obtain the ideal output distribution of a linear optical circuit subject to photon loss [11]. This can be viewed as a form of quantum error mitigation, where the ideal distribution is accessible, but with a sampling cost scaling exponentially with the depth of the circuit. This scaling comes from the fact that the probability of at least one photon being lost approaches one exponentially quickly with increasing circuit depth [12], which results in most outputs being discarded.

Many techniques have been developed to improve the performance of quantum computations run on currently available quantum hardware. These are generally referred to as *quantum error mitigation* (QEM) techniques. Some well-known QEM techniques are zero noise extrapolation (ZNE) [13, 14, 15, 16], probabilistic error cancellation [13, 17, 18, 19], verification-based mitigation [20, 21, 22, 23], virtual distillation and exponential error suppression [24, 25, 26, 27], quantum subspace expansion mitigation [28, 29, 30, 31], and measurement error mitigation [32, 33, 34, 35]

(see [36] for a review of QEM techniques). These techniques are tailored to the circuit model of quantum computing, and thus adapting them to DVLOQC can in some cases be complicated by the considerable differences in the computational setup and the types of noise.

In this work we address the question of whether the effects of photon loss in the output distributions of linear optical quantum circuits can be mitigated by classical postprocessing of lossy output statistics. Crucially, we require that these postprocessing techniques outperform postselection, in the sense that they provide more converged estimates of the ideal probabilities for a comparable sample number. We present various new techniques to mitigate photon loss in linear optical circuits, and perform rigorous analysis of their performance. We refer to these techniques collectively as *recycling mitigation*, as they all involve the use of lossy output statistics that otherwise would be discarded. At the heart of recycling mitigation is the construction of the so-called *recycled probabilities*. The recycled probabilities can be thought of as precursors of the mitigated outputs. Mitigated probabilities are approximations of the ideal probabilities, and can be obtained from the recycled probabilities by appropriate classical postprocessing. We introduce several techniques by which this classical postprocessing step can be performed. These give rise to different photon loss mitigation techniques. Fig. 1 shows, at a very high level, the main steps underlying our mitigation techniques. We provide analytical and numerical evidence that there exists a threshold value, lower bounded by a constant, for the loss per mode η, denoted as η_{th}, such that when $\eta \geq \eta_{th}$, recycling mitigation outperforms postselection.

Furthermore, we provide analytic and numerical evidence showing that mitigation techniques based on artificially increasing noise and Richardson extrapolation, called zero-noise extrapolation (ZNE) [13, 37], such as those applied to mitigate photon loss in the continuous variable regime [38], do not outperform postselection for the problem of mitigating photon loss in the DVLOQC setting. This result is particularly interesting because we also show that, in the case of photon loss, ZNE produces unbiased estimators of the ideal probability. Indeed, we show that the application of ZNE to mitigate photon loss reduces to the problem of inverting a Vandermonde matrix [39] which, in the absence of statistical error, perfectly computes the ideal probabilities. In the presence of statistical error, however, we show that the resulting error on the inversion process is higher than the statistical error of postselection. The requirement of outperforming postselection, which is itself an unbiased estimator of the ideal probability, is what motivated us to search for biased loss mitigation techniques with smaller combined bias and statistical error than postselection.

Recycling mitigation is a biased error mitigation technique, meaning that in addition to statistical error there is also a bias error which, contrary to statistical error, does not decrease with increasing sample size. Nevertheless, we provide strong evidence that: (1) For $\eta \geq \eta_{th}$, there is a sample size up to which the combined statistical and bias errors of recycling mitigation are less than the statistical error of postselection, meaning that recycling mitigation outperforms postselection up to this sample size. (2) We present analytic and numerical evidence that the sample size after which postselection starts to outperform recycling mitigation seems very large, i.e. of the order of $\binom{m}{n}^2$, where $\binom{m}{n}$ is the size of the Fock space. (3) We provide analytic and numerical evidence that, for fixed m and n, and generic interferometers, the bias error in some of our introduced techniques seems small enough such that, in the limit of very low statistical error the mitigated outputs, like the ideal outputs, seem hard to approximate efficiently classically (see Section 6).

Any loss parameter $0 < \eta < 1$ results in a binomial distribution over k, the number of lost photons at the output. The key observation underpinning the utility of recycling is that when $\eta \geq \eta_{th}$, the statistical error denoted ϵ_k^{stat} is greatest for the postselection estimators where $k = 0$, and furthermore that $\epsilon_{k=0}^{stat} \geq \epsilon_{k=1}^{stat} \geq \ldots \geq \epsilon_{k=n_d}^{stat}$, where $k \in \{1, \ldots, n_d\}$ denotes the number of photons lost. ϵ_k^{stat} is the statistical error of recycled probabilities constructed from statistics with k lost photons. In order to obtain the mitigated outputs these techniques use sampled outputs for which $k > 0$ meaning these lossy estimators have lower statistical error than the postselection estimators, but where k is still low enough such that the estimators contain information on the ideal outputs. The fact that statistical errors in recycling mitigation are lower than those of postselection, coupled with the low bias error in our techniques, is what allows the combined bias and statisical errors of the mitigated outputs to be lower than the statisical error of postselection, and thus for our methods to outperform postselection.

Since recycling mitigation is applied to the statistics relating to relatively few lost photons, we

show that the number of samples (runs of the lossy DVLOQC circuit) required for constructing a recycled probability at a given k, and consequently computing the mitigated values to a fixed accuracy, scales exponentially with system size as $O\Big(\frac{1}{\binom{n}{k}(1-\eta)^{n-k}\eta^k}\Big)$. On the one hand, this is a significant improvement over the scaling of postselection when $\eta \geq \eta_{th}$, which is typically $O\big(\frac{1}{(1-\eta)^n}\big)$. On the other hand, exponential scaling is a bottleneck common to all error mitigation protocols involving solely classical postprocessing [40, 41, 42, 43], and not one specific to our technique. Finally, the fact that error mitigation techniques have exponential scaling does not necessarily hinder the achievement of a quantum computational advantage or utility for a fixed system size [44]. Indeed, an initial experimental attempt to achieve a quantum computational utility for a pre-fault tolerant quantum device coupled with error mitigation was presented in [45]. This experiment was later shown to be efficiently simulable classically, albeit using a highly non-trivial tensor network approach [46]. We are hopeful that other attempts in this direction will soon appear. Notably, in the DVLOQC framework where the methods we present can potentially aid in reaching a quantum computational utility in the pre-fault tolerant era.

The photon loss mitigation techniques we introduce can be used in a variety of applications. The work presented in [47] introduced a quantum circuit Born machine (QCBM) [48] tailored to the DVLOQC framework. Performance during the training of the QCBM, when afflicted by photon loss, considerably improved when the recycling mitigation method was applied. Use of the mitigation in the QCBM experiments reduced the overall number of samples needed to reach a desired precision, thereby decreasing the overall computational time. In addition to the application mentioned in [47], other applications include variational quantum eigensolvers [49, 50], photonic differential equation solving [51], photonic quantum machine learning [52], and graph problems with DVLOQC [53].

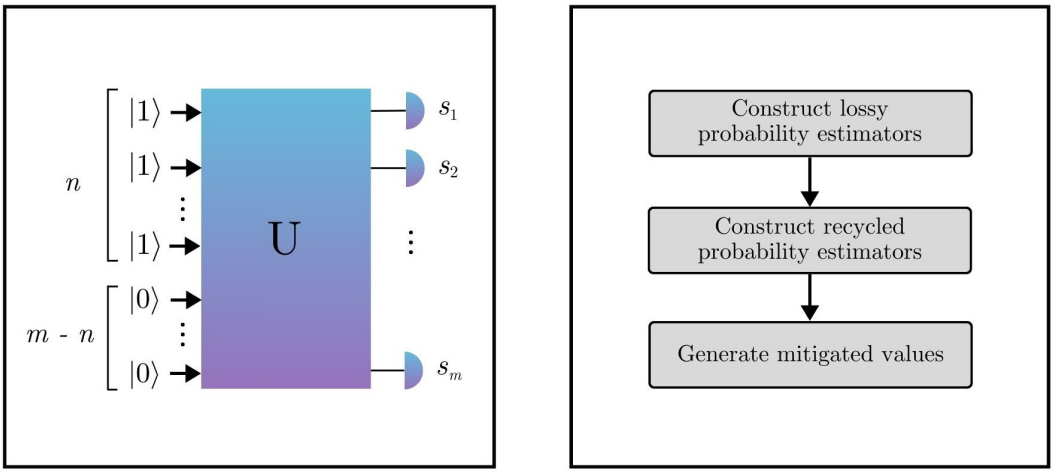
Recent work [54] has provided a method for mitigating photon loss in the continuous variable (CV) setting, adapting pre-existing methods of probabilistic error cancellation [13, 17, 18, 19] to the CV setting. Although the techniques of [54] are applicable to the DVLOQC setting, these are solely for expectation value mitigation (sometimes called weak mitigation) [42]. Our methods, by contrast, can perform both strong (full probability distribution mitigation) as well as weak mitigation, with a classical memory cost scaling with the size of the set of probabilities to be mitigated. Furthermore, we present analytical and extensive numerical evidence that our methods provably outperform postselection, whereas, to our knowledge, no comparison between the performance of the mitigation methods developed and postselection is presented in [54].

This paper is structured as follows. Section 2 gives a high-level overview of our main contributions. Section 3 introduces some notation and basic concepts. Sections 4.1-4.5 detail the construction of the recycled distributions, the classical postprocessing techniques needed to obtain the mitigated distribution. Section 5 contains numerical simulations that aid in understanding how to use recycling mitigation in practice, as well as examples of the techniques in action. Section 6 discusses the properties of the mitigated probabilities in the limit of zero statistical error. Section 7 provides strong evidence that techniques based on ZNE do not in general outperform postselection. Section 8 contains a discussion of our results as well as a set of interesting open questions.

2 Overview of main results

This section summarises the main technical contributions of this work.

First, we compute the overall error on the output probabilities when applying the recycling mitigation technique. That is, we compute the difference between the mitigated and ideal (lossless) output probabilities. This overall error can mainly be seen as a sum of statistical error, as well as the bias error of the technique. This result is stated precisely in the following theorem.



(a) (b)

Figure 1: A schematic illustrating the main steps of the recycling mitigation protocol. (a) An input state of n photons is introduced to a lossy m-mode linear optical interferometer implementing a unitary transformation. The classical measurement outcomes of each of the output modes is denoted by a classical $m-$bit string. (b) The classical data set generated by repeatedly sampling from the quantum circuit is then used as input for classical postprocessing. This classical postprocessing consists of three stages. First, the data set is used to generate lossy probability estimators. These estimators are then used to construct recycled probability estimators, which are in turn then used to generate mitigated values.

Theorem 1. *Consider N_{tot} samples generated from a DVLOQC circuit with n photons and m modes. Let $U \in \mathsf{U}(m)$ be the unitary implemented by the circuit , $\eta \in [0,1]$ the uniform probability a photon is lost in any mode. Assume we are in the no-collision regime, $m \in \Omega(n^2)$, where at most one photon occupies each mode. There exists a classical algorithm (linear solving recycling mitigation) that uses a subset of*

$$N_{rec} = \binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot} \in O\Big(\sqrt{\frac{n}{k(n-k)}}\,\frac{n^n}{k^k(n-k)^{n-k}}\eta^k(1-\eta)^{n-k}N_{tot}\Big),$$

samples where $k > 0$ photons were lost, runs in time $N_{rec} + \mathsf{poly}(m,n,k)$, and that outputs with probability $1-\delta$ an approximation $p_{mit}(s)$ of the exact lossless circuit probability of bit string s, $p_{id}(s)$ such that

$$|p_{mit}(s) - p_{id}(s)| \leq f(n,m,k,\eta,N_{tot}), \tag{1}$$

where $f(n,m,k,\eta,N_{tot}) = 3\left[\sqrt{\frac{2}{\delta}}\left(\frac{e(m-n+k)}{k}\right)^k (\frac{n}{m})^{n/2} + \sqrt{\ln(4/\delta)}\,(\frac{k}{n})^{k/2}\sqrt{\frac{1}{(1-\eta)^{n-k}\eta^k N_{\mathrm{tot}}}}\right]$.

This theorem is proved in Appendix 12.2. The proof makes use of Theorem 10 to bound the bias error of the mitigated output, and as the derived bound does not have an explicit dependence on k it seems likely this could be considerably tightened. As such, Theorem 1 provides a loose upper bound on f. For example, if one looks at minimising f by varying k, one finds $k_{min} = \mathsf{ceil}(\eta n)$ minimises f. It is tempting to say that this is an optimal value of k which should be used in the protocol. However, this is a misleading conclusion directly due to the upper bound for f being loose. Indeed, numerically we find that the error mitigated probabilities for values of k near k_{min} are far from the ideal probabilities. Intuitively, this is because the recycled probabilities used to extract the error mitigated probabilities can efficiently be estimated classically at $k = k_{min}$ by collecting samples from a classical algorithm simulating lossy boson sampling. Indeed, collecting samples from a lossy boson sampler near the average number of lost photons, $\mathsf{ceil}(\eta n)$, can be done efficiently classically for all constant η [12]. Numerically, we find that values of $k \in O(1)$ produce the most accurate mitigation results (see Section 5). Note that these are rare sampling events for the classical sampler of [12]. We believe that any hope of proving the optimality of $k \in O(1)$

values requires a significant tightening of the upper bound on f, which in turn requires a more detailed understanding of the distribution of permanents of random matrices, which to the best of our knowledge remains incomplete [55].

Despite its limitations, Theorem 1 is still useful to show that a regime of advantage over postselection exists, as we will now show.

Postselection uses a subset $N_{post} = (1-\eta)^n N_{tot}$ of samples where $k = 0$ (i.e. no loss occurred) and outputs an estimate $p_{post}(s)$ of $p_{id}(s)$ such that

$$|p_{post}(s) - p_{id}(s)| \leq \gamma \sqrt{\frac{1}{(1-\eta)^n N_{tot}}}, \tag{2}$$

with probability $1 - 2e^{-\gamma^2}$ and runtime $N_{post} + 1$. This upper bound on the additive error is computed via Hoeffding's inequality [56].

The regime of advantage for the mitigation technique over postselection occurs where the combined statistical and bias errors of the mitigated output are lower than the statistical errors of the postselected output. To see the sampling condition where the recycling mitigation outperforms postselection indicated by the analytical bounds one can, in the worst-case where the additive errors of recycling mitigation and postselection equal their upper bounds in eqn. (1) and (2), solve the inequality: $|p_{mit}(s) - p_{id}(s)| \leq |p_{post}(s) - p_{id}(s)|$. This is stated in the following corollary.

Corollary 2. *We first assume that we are operating in the worst-case case error regime where* $|p_{mit}(s) - p_{id}(s)| = f(m, n, k, \eta, N_{tot})$ *(c.f. Theorem 1) and* $|p_{post} - p_{id}| = \gamma\sqrt{\frac{1}{(1-\eta)^n N_{tot}}}$. *Furthermore, we assume that the statistical error of the postselected outputs is higher than the statistical error of the* $(n-k)$*-photon lossy outputs. Recycling mitigation outperforms postselection when*

$$N_{\text{tot}} \leq \frac{\delta(\Delta\eta)^2}{18}\left(\frac{em}{n}\right)^n \left(\frac{k}{e(m-n+k)}\right)^{2k}, \tag{3}$$

where

$$\Delta\eta := \gamma\sqrt{\frac{1}{(1-\eta)^n}} - 3\sqrt{\ln\left(\frac{4}{\delta}\right)}\left(\frac{k}{n}\right)^{k/2}\sqrt{\frac{1}{(1-\eta)^{n-k}\eta^k}}, \tag{4}$$

and, from a union bound, this upper bound on N_{tot} *applies with confidence* $1-\delta'$*, where* $\delta' = \delta + 2e^{-\gamma^2}$.

Note that the assumption in the above corollary that the statistical error from postselection is higher than the statistical error of the mitigated output may be concisely stated as $\Delta\eta > 0$. This corollary means that, in worst-case, there exists a non-trivial sampling regime up to which recycling mitigation outperforms postselection. Furthermore, the condition $\Delta\eta > 0$ indicates that there is a photon loss rate threshold that needs to be satisfied in order for recycling mitigation to outperform postselection. Indeed, $\Delta\eta > 0$ implies the existence of a lower bound on the photon loss rate for which it is possible for mitigation to outperform postselection: $\eta > \eta_{th}(n,k)$ with

$$\eta_{th}(n,k) \geq \frac{1}{1+\frac{en}{k}}.$$

For sufficiently large n and for a fixed k independent of n, we have that $\left(\frac{e(m-n+k)}{k}\right)^{2k} \in O\big(\mathsf{poly}(m,n)\big)$ and $\Delta\eta \in O\left(\frac{1}{\sqrt{(1-\eta)^n}}\right)$. This indicates a regime of advantage for recycling mitigation over postselection up to a sample number

$$N_{tot} \in O\left(\frac{\left(\frac{em}{n}\right)^n}{(1-\eta)^n \mathsf{poly}(m,n)}\right) \in e^{n\log(\mathsf{poly}(n))+O(n)-\log(\mathsf{poly}(n))+O(1)}. \tag{5}$$

Consequently, recycling mitigation outperforms postselection for approximating the ideal probability up to an additive error

$$\epsilon \in O\left(\mathsf{poly}(m,n)\left(\frac{n}{m}\right)^{n/2}\right) \in e^{-\frac{n}{2}\log(\mathsf{poly}(n))+\log(\mathsf{poly}(n))+O(1)}. \tag{6}$$

Intuitively, because the statistical error for recycling mitigation is lower than that for postselection, the point up to which recycling mitigation outperforms postselection is typically determined by the bias error of the mitigation. Indeed, $O(\mathsf{poly}(m,n)\left(\frac{n}{m}\right)^{n/2})$ is a high-confidence upper bound on the bias error obtained in Theorem 10.

We present another recycling mitigation technique, exponential extrapolation, which outperforms the linear solving technique in numerical simulation experiments. Unlike linear solving, which uses a single value of k, exponential extrapolation uses recycled probabilities computed at multiple values of k, and attempts to extract ideal probabilities from the decay of these probabilities towards the uniform distribution (obtained at $k = n$ when all photons are lost). This decay is fitted to an exponential curve. The choice of the decay function is primarily heuristic and accurately captures the form of the decay function observed numerically.

We numerically observe that exponential extrapolation has lower bias error than linear solving. Although we do not show this analytically, we make the following conjecture about the performance of exponential extrapolation based on numerical evidence presented in Section 6.

Conjecture 3. *When $m \in \Omega(n^2)$, and for values of $k \in O(1)$, exponential extrapolation recycling mitigation gives the following performance guarantee with high probability over the choice of U*

$$E_s(\epsilon_{mitbias,s}) \leq \delta(n)\left(\frac{n}{m}\right)^n,$$

where $E_s(\epsilon_{mitbias,s})$ is the average, for fixed U, over all bit strings s of the bias error $\epsilon_{mitbias,s}$, and $\forall n\ \delta(n) < 1$ is some function of n .

By using a Markov inequality, Conjecture 3 being true implies that, with high probability over the choice of U, and with probability $1 - \frac{1}{w}$ over the choice of s for a fixed U and $w > 1$, we have $\epsilon_{mitbias,s} \leq w\delta(n)\left(\frac{n}{m}\right)^n$. In particular, one can choose a w such that $w\delta(n) < 1$, and relabel $\kappa(n) = w\delta(n) < 1$, Conjecture 3 being true therefore implies that with high probability over U and with probability $1 - \frac{1}{w}$ over the choice of s for a fixed U, $\epsilon_{mitbias,s} \leq \kappa(n)\left(\frac{n}{m}\right)^n$ with $\kappa(n) < 1$.

Furthermore, when η is above some threshold value, we numerically find that the statistical error of extrapolation is lower than that of postselection. Let $\epsilon_{stat}(m,n,k,\eta,N_{tot})$ be the upper bound on the absolute value of the statistical error of extrapolation mitigation, we conjecture based on numerical evidence the following.

Conjecture 4. *For $\eta > \eta_{th}$, where $\eta_{th} \in [0,1]$ is some threshold loss value, we have that*

$$\epsilon_{stat}(m,n,k,\eta,N_{tot}) \leq \frac{1}{\sqrt{(1-\eta)^n N_{tot}}}.$$

Furthermore, for large enough n, $\epsilon_{stat}(m,n,k,\eta,N_{tot}) \in o\left(\frac{1}{\sqrt{(1-\eta)^n N_{tot}}}\right)$.

If our conjectures on the bias and statistical errors of exponential extrapolation are true, we can show the following theorem in worst-case, when the errors on postselection and exponential extrapolation equal their upper bounds.

Theorem 5. *Assume that Conjectures 3 and 4 are true. Furthermore, assume that n is sufficiently large and we are in the case where $|p_{mit}(s) - p_{id}(s)| = \epsilon_{stat}(m,n,k,\eta,N_{tot}) + \kappa(n)\left(\frac{n}{m}\right)^n$, with $\kappa(n) < 1$, and $|p_{post}(s) - p_{id}(s)| = \sqrt{\frac{1}{(1-\eta)^n N_{tot}}}$, where $p_{mit}(s)$ is the output probability of exponential extrapolation recycling mitigation. Then exponential extrapolation outperforms postselection for up to sample number $N_{tot} \in O\left(\frac{(\frac{em}{n})^{2n}}{\kappa(n)^2(1-\eta)^n}\right) \in e^{2n\log(\mathsf{poly}(n))+O(n)-\log(\kappa(n))+O(1)}$ and up to additive error $\epsilon \in O\left(\kappa(n)\left(\frac{n}{m}\right)^n\right) \in e^{-n\log(\mathsf{poly}(n))+\log(\kappa(n))+O(1)}$.*

A proof of this is provided in Appendix 13.2.

Theorem 5 indicates that in worst-case, exponential extrapolation has better performance than linear solving, as the additive error up to which exponential extrapolation outperforms postselection is significantly smaller than that of linear solving.

In Section 5, we present the results of numerical simulations comparing the performance of the mitigation techniques with postselection. The recycling mitigation methods (linear solving and exponential extrapolation) consistently outperform postselection for a range of loss values and sample numbers. We also present results that indicates that in the average-case the errors of both recycling mitigation methods and postselection are below their analytically computed upper bounds. These experiments were performed in loss regimes relevant to near-term quantum photonics hardware.

In Section 6, we analyse the mitigated outputs in the regime $m \in \Omega(n^2)$ for exponential extrapolation error mitigation, and conjecture that, in the absence of statistical errors, the distribution of $p_{mit}(s)$ is likely classically hard to compute. We support this conjecture with numerical calculations. Specifically, the conjecture is that computing the mitigated outputs is a problem that reduces to solving a restricted version of the $|GPE_{\pm}|^2$ problem [6]. This result is interesting because it indicates that, in general and in the absence of statistical error, the bias errors are small enough to render the error mitigated probabilities hard to compute classically.

Our final contribution is in providing strong evidence that photon loss error mitigation techniques based on ZNE [36] cannot outperform postselection in the DVLOQC setting. Indeed, in Section 7 we provide an upper bound on the error incurred from ZNE, E_{extrap}, which is larger than the worst-case statistical error of postselection (see Thm. 16). In Appendix 15, we also provide numerical evidence that for all $n \geq n_0$, for some $n_0 \in \mathbb{N}$, E_{extrap} is always larger than the statistical error of postselection. Indeed, in DVLOQC there is a natural way to mitigate photon loss errors, postselection, which is not present for other types of errors in other types of hardware [36]. This therefore sets a benchmark that a photon loss mitigation technique in DVLOQC must satisfy to be useful, namely that it must outperform postselection.

3 Preliminaries

Consider the DVLOQC setting where a photonic quantum device is composed of a single-photon source [7], a universal m-mode linear optical interferometer [8] capable of implementing any unitary transformation $U \in \mathsf{U}(m)$, with $\mathsf{U}(m)$ the group of unitary $m \times m$ matrices, and single photon detectors [9]. Generating a sample using this device proceeds as follows. First, n single photons emitted from the source pass through the linear optical inteferometer in an input configuration $\mathbf{T} := (t_1, \ldots, t_m)$, where t_i is the number of photons in mode i. Let $|\psi_{in}\rangle := |t_1, \ldots, t_m\rangle$ be the input Fock state of single-photons corresponding to the configuration $\mathbf{T}$. The linear optical interferometer implements a unitary transformation $\phi(U)$ on the input state $|\psi_{in}\rangle$ [6], resulting in an output state $|\psi_{out}\rangle := \phi(U)|\psi_{in}\rangle$, where $\phi(U)$ represents the action of the unitary U, implemented by the interferometer, on $|\psi_{in}\rangle$. Note that $\phi(U)$ and U are related by a homomorphism, detailed in [6]. A sample $\mathbf{S} := (s_1, \ldots, s_m)$, with s_i the number of photons in output mode i, is then obtained by measuring the number of photons in each output mode using single-photon detectors. This corresponds to projecting $|\psi_{out}\rangle$ onto the state $|s_1, \ldots, s_m\rangle$. Many computational tasks on photonic quantum devices can be implemented by collecting samples according to the previous procedure, and then performing classical postprocessing [49, 53, 57, 50, 1].

In the absence of any errors affecting the device, $\sum_{i=1,\ldots,m} t_i = \sum_{i=1,\ldots,m} s_i = n$, furthermore, the probability of obtaining the sample $\mathbf{S}$ is proportional to the modulus squared of the permanent of a submatrix $U_{\mathbf{T},\mathbf{S}}$ of U, whose rows and columns are determined by the input and output occupancies $\mathbf{T}$ and $\mathbf{S}$ [6]. By appropriately choosing the unitary transformation U, and performing the above mentioned sampling procedure repeatedly, one can perform both non-universal and universal quantum computing with linear optics. In particular, if U is chosen to be Haar random, one performs boson sampling, a non-universal sampling task which is hard for classical computers to carry out efficiently [6]. Alternatively, choosing specific unitaries U, and postselecting on detecting a specific output configuration, one can perform universal quantum computation [1].

We will now describe our error model as well as the assumptions we will make throughout this paper.

- We consider photon loss as the only source of error affecting our devices. Our error model is the uniform loss model, where a photon is equally likely to be lost in any mode $i \in \{1, \ldots, m\}$ with probability $\eta \in [0, 1]$. Following the commutation rules of photon loss [58], we assume

without loss of generality that the photons are lost at the output of the interferometer, just before the single-photon detectors which are assumed to be perfect.

- We assume that sampling occurs in the no-collision regime, where at most one photon occupies any output mode. This is approximately true for $m \in \Omega(n^2)$ [6]. In this regime the samples are given by $\mathbf{S} = (s_1, \ldots, s_m)$, with $s_i \in \{0, 1\}$, and are therefore bit strings of length m. The total number of possible no-collision outputs is $\binom{m}{n}$.

The uniform loss model is a widely used error model for simulating photon loss, and is the standard assumption used when constructing loss-tolerant quantum error correcting codes [59]. It is also often assumed when deriving efficient classical algorithms for simulating lossy linear optical setups [60]. Working in the no-collision regime is primarily interesting for two reasons. Firstly, one can prove statements of quantum advantage in boson sampling in this regime [6]. And secondly, the fact that at most one photon occupies each mode allows us to assume the use of the standard and widely available threshold detectors, rather than number resolving ones which are currently challenging to practically implement. As a final note, while the no-collision assumption is a useful one, it is not necessary for our techniques to work. Indeed, as discussed in later parts of this paper, our techniques can in principle be generalised to the case where more than one photon can occupy a mode.

The uniform loss model induces a binomial distribution on the samples, in the sense that sampling N_{tot} times from a uniformly lossy linear optical circuit produces approximately $N_{tot,k} := \binom{n}{k} N_{tot} \eta^k (1-\eta)^{n-k}$ samples corresponding to k lost photons, for $k \in \{0, \ldots, n\}$. Note that $N_{tot} = \sum_{k=0,\ldots,n} N_{tot,k}$. We will take s_i^{n-k} to mean a bit string of the form $\{s_1, \ldots, s_m\}$ where $\sum_i s_i = n - k$, $s_i \in \{0, 1\}$ is the number of photons in mode i, and $k \in \{0, \ldots, n\}$. This corresponds to a sample drawn from the probability distribution where k of the initial n input photons have been lost. In order to estimate the probability $p(s_i^{n-k})$ from a set $\mathcal{W} := \{s_j^{n-k}\}$ of samples [1] where $|\mathcal{W}| \leq N_{tot,k}$, we perform the following procedure. For each w ranging from 1 to $|\mathcal{W}|$, assign a value 1 to a random variable $X_w \in \{0, 1\}$ if the sample s_w^{n-k} is the bit string s_i^{n-k}, and assign the value 0 to X_w otherwise. The estimate $\tilde{p}(s_i^{n-k})$ is then

$$\tilde{p}(s_i^{n-k}) := \frac{\sum_w X_w}{|\mathcal{W}|}. \tag{7}$$

This estimation therefore induces a statistical error given by

$$\epsilon_{stat}(s_i^{n-k}) := |\tilde{p}(s_i^{n-k}) - p(s_i^{n-k})|. \tag{8}$$

In postselection, estimators are constructed using only the outputs for which $k = 0$, when photon loss is uniform this corresponds to approximately $N_{tot}(1-\eta)^n$ samples. As the the size of the system is increased, the probability of postselecting non-lossy outcomes decays exponentially towards zero.

For values of loss above a certain threshold, the statistical error of probabilities constructed from lossy statistics is in general lower than those constructed from lossless ones. As an example of how to see this, note that for $k = 1$, the expected number of samples is $N_{tot} n (1-\eta)^{n-1} \eta$. For the lossless $k = 0$ case, we have $N_{tot}(1-\eta)^n$ such samples. Since the statistical error is typically upper bounded by $1/\sqrt{N}$, where N is the sample number, we can see that the inequality $N_{tot} n (1-\eta)^{n-1} \eta \geq N_{tot}(1-\eta)^n$ (which implies lower statistical errors for lossy probabilities) holds whenever $\eta \geq \frac{1}{n+1}$.

Recycling mitigation uses the $n-k$–photon probability estimates $\{\tilde{p}(s_i^{n-k})\}$, potentially for a range of k values, and construct from these a *mitigated* n–photon probability distribution $\{p_{mit}(s_j^n)\}$. To have any utility, a photon loss mitigation technique needs to outperform computing the n–photon probability estimates $\{\tilde{p}(s_j^n)\}$ from the samples $N_{tot,0}$, which we will henceforth refer to as postselection on n–photon outputs, or just postselection. We will therefore use postselection as the benchmark to evaluate the performance of recycling mitigation. Postselection is, to our knowledge, the only technique being used to mitigate the effects of photon loss on current DVLOQC

[1] The set can contain repeated identical bit strings, i.e. there can be $j_1 \neq j_2$, such that $s_{j_1}^{n-k} = s_{j_2}^{n-k}$.

hardware [49]. Another factor motivating recycling mitigation is that it does not increase the overall sample cost relative to postselection. This contrasts favourably with many error mitigation results that have an accompanying sampling overhead [36].

4 Method

4.1 Recycled probabilities

The recycled probabilities are constructed from $n-k$ output photon statistics, where $k \in \{1,\ldots,n-1\}$. To explicitly analyse the signal of the ideal probability within the recycled probability, the recycled probabilities may be decomposed into a combination of an ideal n–photon output probability, and an interference term consisting of a mixture of other n–photon output probabilities from the distribution. We first describe the construction of the recycled probabilities from $n-k$ output photon statistics, which generalises for all k. We then describe the analytical decomposition of the recycled probabilities into n–photon output probabilities.

4.1.1 Construction of recycled probabilities from lossy outputs

We now detail the construction of recycled probabilities from $n-k$–photon output statistics - that is, the output statistics in which exactly k of n photons have been lost. This construction should be applied to obtain the recycled probability distribution in an experiment. The recycled probability for bit string s_l^n computed from $n-k$–photon output statistics is denoted $p_R^k(s_l^n)$, with $k \in \{1,\ldots,n-1\}$. Hence there are $n-1$ recycled probabilities one can construct from lossy output statistics for any n–photon output bit string, one for each possible value of k. Performing the construction involves summing over the $n-k$ output photon bit string probabilities that relate to a particular ideal probability. Informally, this relation is that these are the probabilities of $n-k$–photon output bit strings that the ideal output bit string can be mapped to through the loss of k photons. We now provide a formal statement of this relation.

We first define a mapping procedure from each n–photon output bit string to a set of $n-k$–photon output bit strings. Each mapped set of bit strings represents the set of all possible states that the associated n–photon output state could become after losing k photons. Let $\mathcal{M}_{\text{unocc.},k,i}$ be the subset of $\{1,\ldots,m\}$ corresponding to the unoccupied modes of the output bit string s_i^{n-k}. That is, the set of indices of the modes $j \in \{1,\ldots,m\}$ of the bit string for which $s_j = 0$. The number of unoccupied modes for $n-k$–photon outputs is $|\mathcal{M}_{\text{unocc.},k,i}| = m-n+k$. Let $\overline{\mathcal{M}_{\text{unocc.},k,i}}$ be the complement of $\mathcal{M}_{\text{unocc.},k,i}$ in $\{1,\ldots,m\}$, so that $\overline{\mathcal{M}_{\text{unocc.},k,i}} \cup \mathcal{M}_{\text{unocc.},k,i} = \{1,\ldots,m\}$. The subset $\overline{\mathcal{M}_{\text{unocc.},k,i}}$ denotes the occupied modes of s_i^{n-k}, consisting of the set of indices of modes $j \in \{1,\ldots,m\}$ for which $s_j = 1$. As the bit string s_i^{n-k} represents an $n-k$–photon output the number of occupied modes is $|\overline{\mathcal{M}_{\text{unocc.},k,i}}| = n-k$. We define the set $\mathcal{L}(s_i^n) := \{s_j^{n-k} | s_i^n \Rightarrow s_j^{n-k}\}$. And the set of all size k subsets of $\overline{\mathcal{M}_{\text{unocc.},0,i}}$ is $\overline{\mathcal{S}_{0,i}} := \{X \subset \overline{\mathcal{M}_{\text{unocc.},0,i}} \,|\, |X| = k\}$. The symbol '$\Rightarrow$' denotes the operation where, for every size k subset $\{\overline{l_1},\ldots,\overline{l_k}\} \in \overline{\mathcal{S}_{0,i}}$, the n–photon output bit string s_i^n is mapped to a new $n-k$–photon output bit string s_j^{n-k} by replacing $s_{\overline{l_i}} = 1$ with $s_{\overline{l_i}} = 0$. In this case, the number of $k-$subsets of $\overline{\mathcal{M}_{\text{unocc.},0,i}}$ is $\binom{n}{k}$, and so $|\mathcal{S}_{0,i}| = \binom{n}{k}$ and the size of the generated set of bit strings is $|\mathcal{L}(s_i^n)| = \binom{n}{k}$.

The recycled probability for the bit string s_l^n can be defined as the sum of probabilities of the $\binom{n}{k}$ bit strings $s_i^{n-k} \in \mathcal{L}(s_l^n)$,

$$p_R^k(s_l^n) := \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} p(s_i^{n-k}). \tag{9}$$

To ensure the normalisation of the recycled distribution the above expression is multiplied by a normalisation factor $\mathbf{N} = \frac{1}{\binom{m-n+k}{k}}$, so that in practice it is

$$p_R^k(s_l^n) = \frac{1}{\binom{m-n+k}{k}} \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} p(s_i^{n-k}). \tag{10}$$

A derivation of the normalisation factor is provided in the next section.

To illustrate how this construction might work in practice we now provide a small example. If we would like to compute the recycled probability for the bit string: 111000, in an experiment in which there are $m = 6$ modes, $n = 3$ input photons and the construction is being performed for $k = 1$ lost photons. The recycled probability may be computed directly from eqn. 10 as being

$$p_R^1(111000) = \big(p(110000) + p(101000) + p(011000)\big)\binom{4}{1}^{-1}, \tag{11}$$

where the normalising parameter is $\mathbf{N} = \binom{4}{1}^{-1}$.

In practice, rather than using the set of exact probabilities, $\{p(s_i^{n-k})\}$, to compute recycled probabilities in the manner shown in eqn. 10, instead empirical estimates of the exact probabilities, $\{\tilde{p}(s_i^{n-k})\}$, are used. These are calculated from the set of measured experimental output bit strings, as described for eqn. 7, and so include statistical errors due to finite samples. This statistical error is an important consideration when making comparisons with postselection, and is later included in the analysis of the protocol performance. We now provide pseudocode detailing how to construct the recycled probability for a specific n output photon bit string and value of k from a set of N output bit strings sampled from a given circuit.

Algorithm 1: Construction of a recycled probability estimator from output statistics

input : The n–photon output bit string s_l^n for which the recycled probability estimator is to be constructed, a set of N output sample bit strings from the DVLOQC circuit $\{s_j\}_{j\in\{1,\ldots,N\}}$, and the choice of k value indicating that $n-k$ output photon statistics be used for the construction.

Initialise variable $\tilde{p}_R^k(s_l^n) \leftarrow 0$ for the recycled probability estimator to be computed.
Create a new set of bit strings by discarding all output bit strings from set $\{s_j\}_{j\in\{1,\ldots,N\}}$ except those for which the number of measured output photons was $n-k$, with the new list denoted $\{s_l\}_{l\in\{1,\ldots,N_{est,k}\}}$ where $N_{est,k} \leq N$.
Generate the set of $n-k$–photon lossy output bit strings $\mathcal{L}(s_l^n)$.
Initialise variable $X_{s_l} \leftarrow 0$.
for $\mathrm{l} = 1$ **to** $N_{est,k}$ **do**
    if $s_l \in \mathcal{L}(s_l^n)$ **then**
        $X_{s_l} \leftarrow X_{s_l} + 1$.
    end
end
Update recycled probability estimator variable as $\tilde{p}_R^k(s_l^n) \leftarrow \frac{X_{s_l}}{\binom{m-n+k}{k} N_{est,k}}$.

output: Recycled probability estimator $\tilde{p}_R^k(s_l^n)$

4.1.2 Decomposition of recycled probabilities into n–photon output bit string probabilities

In later sections, the recycled probabilities are analysed in terms of their decomposition into n–photon output probabilities. For a given output bit string s_l^n, this representation allows explicit treatment of the signal of the ideal n–photon output probability $p(s_l^n)$ within the recycled probability $p_R^k(s_l^n)$. To get the recycled probabilities in this form, the lossy output probabilities within the sum in eqn. 10 are decomposed into n–photon output probabilities from the ideal distribution. The details of this decomposition will now be formalised.

For this purpose we now define another mapping procedure, this time from each lossy $n-k$–photon output bit string from the sum in eqn. 10 to a set of n–photon output bit strings. Where each mapped set of bit strings represents the set of n–photon outputs that the associated lossy output bit string could have been had loss not occurred. The set of all size k subsets of $\mathcal{M}_{\text{unocc.},k,i}$ is defined $\mathcal{S}_{k,i} := \{X \subset \mathcal{M}_{\text{unocc.},k,i} |\ |X| = k\}$. We define the set $\mathcal{G}(s_i^{n-k}) := \{s_j^n | s_i^{n-k} \rightarrow s_j^n\}$. The symbol '$\rightarrow$' denotes the operation where, for every size k subset $\{l_1, \ldots, l_k\} \in \mathcal{S}_{k,i}$, the bit string s_i^{n-k} is mapped to a new bit string by replacing each bit $s_{l_i} = 0$ with $s_{l_i} = 1$. The number of $k-$subsets of $\mathcal{M}_{\text{unocc.},k,i}$ is $\binom{m-n+k}{k}$, and so $|\mathcal{S}_{k,i}| = \binom{m-n+k}{k}$ and the size of the generated set of bit strings is $|\mathcal{G}(s_i^{n-k})| = \binom{m-n+k}{k}$.

From the addition rule of probabilities and the uniformity of the loss, the probability $p(s_i^{n-k})$ of obtaining the output bit string s_i^{n-k} is

$$p(s_i^{n-k}) = \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) \frac{1}{\binom{n}{k}}, \tag{12}$$

where the sum includes all the bit string outputs $s_j^n \in \mathcal{G}(s_i^{n-k})$ from which the loss of k photons maps to s_i^{n-k}. Each element of the sum is composed of the probability $p(s_j^n)$ that the output is the n–photon bit string s_j^n, multiplied by the uniform probability $\binom{n}{k}^{-1}$ of the loss of k photons from s_j^n resulting in the output s_i^{n-k}.

As in eqn. 12, in the definition of the recycled probabilities in eqn. 10 each of the lossy probabilities can be decomposed into a convex combination of probabilities from the ideal distribution. That is, the $n-k$–photon output probabilities within the sum in eqn. 10 can be replaced with a sum over n–photon output probabilities. For some arbitrary labelling of bit strings s_j^n, and noting that $s_l^n \in \mathcal{G}(s_i^{n-k})$, we can expand the $n-k$ output photon probabilities in the form

$$p(s_i^{n-k}) = p(s_l^n) \frac{1}{\binom{n}{k}} + \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\binom{n}{k}}. \tag{13}$$

This can then be used to separate the contribution of the ideal bit string probability $p(s_l^n)$ out from the rest of the probabilities which are grouped into a sum we call an *interference term* in the recycled probability. So it becomes

$$p_R^k(s_l^n) = p(s_l^n) + \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\binom{n}{k}}. \tag{14}$$

The last step is to normalise the distribution $p_R^k(s_l^n)$, namely to compute $\mathbf{N}$ such that $\mathbf{N} \cdot \sum_l p_R^k(s_l^n) = 1$. To do this, note that

$$\begin{aligned} \sum_{l=1}^{\binom{m}{n}} p_R^k(s_l^n) &= \sum_{l=1}^{\binom{m}{n}} \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} p(s_i^{n-k}) \\ &= \sum_{l=1}^{\binom{m}{n}} \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) \frac{1}{\binom{n}{k}} \\ &= \binom{m-n+k}{k}, \end{aligned} \tag{15}$$

This is because $\sum_l p(s_l^n) = 1$, and, because $|\mathcal{L}(s_i^n)| = \binom{n}{k}$ and $|\mathcal{G}(s_i^{n-k})| = \binom{m-n+k}{k}$, each distinct bit string s_l^n appears exactly $\binom{n}{k}\binom{m-n+k}{k}$ times in the above sum [2]. Therefore $\mathbf{N} = \frac{1}{\binom{m-n+k}{k}}$, and the expression for the normalised recycled probability is

$$p_R^k(s_l^n) = p(s_l^n) \frac{1}{\binom{m-n+k}{k}} + \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\binom{m-n+k}{k}\binom{n}{k}}. \tag{16}$$

Let $N_k := \binom{m-n+k}{k}\binom{n}{k}$, $N_k' := \left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}$. The recycled probability $p_R^k(s_l^n)$ is composed of the ideal output probability $p(s_l^n)$ and an interference term, defined as

$$I_{s_l^n,k} := \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{N_k'}. \tag{17}$$

[2] Another way to see this is to consider the sum over recycled probabilities as $\sum_l p_R(s_l^n) = \sum_l \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} p(s_i^{n-k}) = \binom{m-n+k}{k}$. Where the rightmost equality follows from the fact that each $p(s_i^{n-k})$ appears exactly $\binom{m-n+k}{k}$ times in this sum, since there are $|\mathcal{G}(s_i^{n-k})|$ recycled probabilities sharing a single $p(s_i^{n-k})$. Furthermore note that $\sum_i p(s_i^{n-k}) = 1$.

The recycled probability may then be written explicitly in this form

$$p_R^k(s_l^n) = p(s_l^n)\frac{1}{\binom{m-n+k}{k}} + \frac{N_k'}{N_k} I_{s_l^n,k}. \tag{18}$$

Importantly, the photon recycled probability $p_R^k(s_l^n)$ contains an amplified signal of the ideal probability $p(s_l^n)$, by a factor of $\binom{n}{k}$, relative to the other probabilities contained within the interference term.

4.2 A classical simulation algorithm for recycled probabilities

In this section, we show that recycled probabilities constructed from output statistics where most photons were lost (i.e. when $n-k$ is a constant independent of n) are efficiently classically computable, and therefore not useful for obtaining interesting mitigation performance. Furthermore, we provide evidence that recycled probabilities constructed from $n-k$ output statistics where k is a constant independent of n (corresponding to output statistics where a small number of photons have been lost) are hard to compute classically, making these interesting to use for recycling mitigation protocols.

Recall that, up to normalisation, a recycled probability is a sum of the form

$$\sum_{s_i^{n-k}\in\mathcal{L}(s_l^n)} p(s_i^{n-k}),$$

where the number of terms of this sum is $\binom{n}{k}$. For Haar-random interferometers U, in the no-collision regime where $m=\Omega(n^5)$, we can use the results of [61], and in turn express each $p(s_i^{n-k})$ as

$$p(s_i^{n-k}) = \frac{1}{\binom{n}{k}}\sum_i \frac{|\mathsf{Per}(X_i)|^2}{m^n},$$

where the X_i's are $n-k\times n-k$ matrices with independently distributed Gaussian entries [6], and the number of terms in this sum is also $\binom{n}{k}$. We show that in the high loss regime, where $k=n-r$, and r is a constant independent of n, the sum $\sum_{s_i^{n-k}\in\mathcal{L}(s_l^n)} p(s_i^{n-k})$ is computable efficiently classically. Our result is encompassed in the following lemma proven in appendix 9.

Lemma 6. *Let $k=n-r$, there is a classical algorithm running in time $O(2^{r-1}r\binom{n}{n-r}^2)$ which exactly computes $\sum_{s_i^{n-k}\in\mathcal{L}(s_l^n)} p(s_i^{n-k})$.*

Notice that when r is a constant independent of n, the runtime of the above algorithm of the order of $\binom{n}{n-r}\in O(n^r)$ which is a polynomial in n. Thus, recycled probabilities corresponding to a high number of lost photons are efficiently computable classically. For low values of loss, in particular when r scales with n, the above efficient classical simulability results break down, this however does not necessarily imply that there is no other, possibly efficient, algorithm for computing $\sum_{s_i^{n-k}\in\mathcal{L}(s_l^n)} p(s_i^{n-k})$.

For the case where k is a constant independent of n, there is evidence that computing $\sum_{s_i^{n-k}\in\mathcal{L}(s_l^n)} p(s_i^{n-k})$ is probably hard (inefficient) to do classically (see also the discussion in Section 6). Indeed, it is known that in worst-case the probabilities $p(s_i^{n-k})$ when k is a constant independent of n are not efficient to compute classically, unless the polynomial hierarchy collapses to its third level [61]. This motivates the application of recycling mitigation in the low k regime, as the $n-k$–photon output statistics are efficiently computable classically when k is high, and therefore any potential quantum advantage is lost. In all our simulations, we apply our error mitigation techniques to statistics where the number of lost photons is a constant independent of system size, and discard all other statistics.

4.3 Bounding the bias error

In this section, we give bounds on the bias error induced when replacing the interference term with its expectation value. This bias error will be present in all the error mitigation techniques we

derive later on. We provide probabilistic bounds on this error, using statistical inequalities. Also, we provide a *deterministic* upper bound by using the results of [62] which hold for specific families of unitary matrices.

The recycled probabilities have a composite structure, comprising a mixture of the ideal probability and the interference term. While the interference term,

$$I_{s_l^n,k} = \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{N_k'}, \tag{19}$$

is itself a mixture of n–photon output probabilities. It is possible to account for the errors caused by using approximations of the interference terms when generating the mitigated probabilities by upper bounding the deviation of the interference term away from an expected value.

Firstly, we consider the case of Haar random matrices. As we are working in the no-collision regime, the output probabilities generated by sampling from a Haar random matrix is linked to permanents of Gaussian random matrices $X \in \mathcal{G}_{n\times n}$ [61] with $\mathcal{G}_{n\times n}$ the set of all such Gaussian matrices; more precisely the set of complex matrices whose real and imaginary parts are chosen independently from the normal distribution $\mathcal{N}(0, \frac{1}{2})$. Let $p_{unif} := \frac{1}{\binom{m}{n}}$, in appendix 11 we show the following.

Lemma 7. *For all s_l^n, we have*

$$\mathbf{E}_{X\in\mathcal{G}_{n\times n}}\left(I_{s_l^n,k}\right) = p_{unif}\left(1 + g(m)\right) \approx p_{unif}, \tag{20}$$

where $g(m) \in O(m^{-1})$, with $\mathbf{E}_{X\in\mathcal{G}_{n\times n}}(.)$ the expectation value over the set $\mathcal{G}_{n\times n}$.

With Lemma 7 in hand, it is possible to bound the deviation of the interference terms around this expected value according to the theorem shown in Appendix 11.

Theorem 8. *The deviation of interference terms around p_{unif} for Haar random matrices is bounded*

$$Pr\Big(\left|I_{s_l^n,k} - p_{unif}\right| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{np_{unif}^2}{\epsilon_{bias,s_l^n}^2}, \tag{21}$$

where ϵ_{bias,s_l^n} is a positive real number.

Secondly, as it is desirable not to be restricted to only mitigating the output of Haar random matrices, we derive an additional bound for arbitrary matrices. Let D_R^k be the uniform distribution over recycled probabilities $\{p_R^k(s_l^n)\}_l$. Since every recycled probability has an associated interference term $I_{s_l^n,k}$, one can equivalently think of D_R^k as a distribution over interference terms. A random variable Y chosen from D_R^k means choosing, with uniform probability p_{unif}, a value from the set $\{p_R^k(s_l^n)\}_l$ (or, equivalently, choosing a value from $\{I_{s_l^n,k}\}_l$ uniformly randomly). In Appendix 11, we show the following.

Lemma 9.

$$\mathbf{E}_{D_R^k}\left(I_{s_l^n,k}\right) = p_{unif}, \tag{22}$$

where $\mathbf{E}_{D_R^k}(.)$ denotes the expectation value over D_R^k.

The deviation of interference terms around this expected value is upper bounded according to the following inequality.

Theorem 10. *The deviation of interference terms around p_{unif} for an arbitrary matrix is bounded*

$$Pr\Big(\left|I_{s_l^n,k} - p_{unif}\right| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{p_{unif}}{\epsilon_{bias,s_l^n}^2}, \tag{23}$$

where ϵ_{bias,s_l^n} is a positive real number.

While both these upper bounds use Chebyshev's inequality [63], the Thm. 8 bound applies only to Haar random matrices, while the Thm. 10 bound applies to arbitrary matrices. However, to upper bound the variance for the Haar random case we use exact moments originally derived in [55]. While in the case of arbitrary matrices the Bhatia-Davis inequality [64] is instead used, which results in a looser bound.

We note that a tighter, although distribution-dependent, bound may be found than the one provided in Thm. 10. This result follows from Lemmas 11 and 12, proven in Appendix 11, which will now be stated, that use the definition of the variance that for a set of real values $\{x_i\}_i$ with mean μ and cardinality N, $\mathsf{Var}(\{x_i\}_i) := N^{-1}\sum_i (x_i-\mu)^2$.

Lemma 11. *The variance of the set of recycled probabilities is less than or equal to the variance of the set of ideal probabilities, that is*

$$\mathsf{Var}\big(\{p(s_l^n)\}_l\big) \geq \mathsf{Var}\big(\{p_R^k(s_l^n)\}_l\big). \tag{24}$$

Lemma 12. *The variance of the set of interference terms is less than or equal to the variance of the set of ideal probabilities, that is*

$$\mathsf{Var}\big(\{p(s_l^n)\}_l\big) \geq \mathsf{Var}\big(\{I_{s_l^n,k}\}_l\big). \tag{25}$$

Using these lemmas, an upper bound on the largest probability of the output distribution may be used to derive a tighter upper bound on the variance using the Bhatia-Davis inequality [64]. Let p_{upper} be an experimentally derived upper bound on the largest probability in the ideal n–photon output distribution $p_{\max}$, such that $p_{\text{upper}} \geq p_{\max}$. Following similar steps as for the proof for Thm. 10 results in an upper bound on the confidence of $1 - \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}} - \delta\Big(1 - \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}}\Big)$, where δ is the confidence parameter for the p_{upper} estimator. This result is stated formally in the following.

Theorem 13. *The deviation of interference terms around p_{unif} for an arbitrary matrix is bounded*

$$Pr\Big(\big|I_{s_l^n,k} - p_{unif}\big| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}} + \delta\Big(1 - \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}}\Big),$$

where ϵ_{bias,s_l^n} is a positive real number, and p_{upper} is an empirically computed upper bound on the largest probability of the ideal n output photon probability distribution with confidence $1-\delta$.

A proof of this result is given in appendix 11. We note also that as $p_{\max} \leq 1$ with confidence 1, so that $\delta = 0$ in Thm. 13, Thm. 10 follows as a corollary.

With assumptions about the structure of the unitaries, like in [62], it is possible to deterministically and exponentially upper bound the interference terms.

For a large class of unitary matrices, we show that the bias error scales as an inverse exponential in n. As will be seen later, this shows that our mitigation techniques outperform postselection in estimating output probabilities and expectation values of linear optical circuits for up to inverse exponential precisions. Here we use the the operator 2-norm and the infinity norm. For an $n \times n$ matrix A the operator 2-norm is defined $\|A\|_2 := \sup_{\|\vec{x}\|_2 \leq 1, \vec{x}\in\mathbb{C}^n} \|A\vec{x}\|_2$, where $\|\vec{v}\|_p$ is the l_p norm, that is $\|\vec{v}\|_p = \big(\sum_p |v_i|^p\big)^{1/p}$. And the infinity norm which is defined as $\|\vec{v}\|_\infty := \max_i |v_i|$. Let $h^A_\infty := \frac{1}{n}\sum_{i=1,\ldots,n} \|\mathbf{A}_i\|_\infty$, where $\mathbf{A}_i$ is the ith row of A. In Appendix 14, we use a result of [62] to show the following.

Theorem 14. *For the class of unitary matrices U with submatrices A such that $p_{max} := \mathsf{max}_{s_l^n}(p(s_l^n)) = |\mathsf{Per}(A)|^2$, and where these matrices A satisfy $\frac{h^A_\infty}{\|A\|_2} \ll 1$, the bias error ϵ_{bias,s_l^n} is bounded*

$$\epsilon_{bias,s_l^n} \in O(e^{-2\times 10^{-5} n}). \tag{26}$$

Interestingly, the probabilistic bounds of Theorems 8 and 10 also give, with high-confidence, exponentially low bounds on the bias error. Indeed, by setting for example $\epsilon_{bias,s_l^n} \in \Omega(\sqrt{p_{unif}})$ in Theorem 10, we observe that $|I_{s^n,k} - p_{unif}|$ is upper bounded by $\Omega(\sqrt{p_{unif}})$ with confidence $1-o(1)$. Intuitively, and as will be seen later in more details, both the probabilistic and deterministic upper

bounds on the bias error indicate that recycling mitigation outperforms post-selection for up to inverse-exponential in n precision.

Appendix 16.1 discusses possible directions to improve our performance guarantees, and contains a technical result about sums of permanents of i.i.d. Gaussian matrices [6] that might find use beyond this work.

4.4 Generating the loss-mitigated outputs

We now present two methods for constructing loss mitigated outputs from the recycled probabilities, these we refer to as *linear solving* and *extrapolation*. In linear solving, the interference term within each recycled probability is substituted for its expected value, and the resulting expressions are then solved to find estimators of the ideal probabilities. While in extrapolation, the decay of the ideal signal in the set of recycled distributions with k is used to compute estimators of the ideal probabilities.

While the postprocessing required to generate each mitigated value may be performed efficiently, the size of the output distribution is exponential in m and n. Meaning the classical postprocessing cost (i.e. the memory cost) is proportional to the number of mitigated probabilities that are to be generated. So that to generate a full mitigated output distribution this scales as m^n, while for a subset of size N_s the postprocessing cost is then proportional to N_s.

4.4.1 Linear solving

The linear solving method involves substituting the interference term within each recycled probability for an approximate value, and then solving the resulting expressions for the ideal probabilities. As the expectation of interference terms over the Haar measure and over the recycled distribution are $p_{unif} + O(m^{-n+1})$ and p_{unif}, respectively, see Lemmas 7 and 9, the term $\frac{N'_k}{N_k} p_{unif}$ is used for this substitution. This allows the upper bounding of the error introduced by the substitution using Thms. 8 and 10. Each recycled probability constructed from the $n-k$ output photon statistics may then be written

$$\tilde{p}^k_R(s^n_l) = \frac{p(s^n_l)}{\binom{m-n+k}{k}} + \frac{N'_k}{N_k}\Big(p_{unif} + \epsilon_{bias,s^n_l}\Big) + \epsilon_{stat.,s^n_l}, \tag{27}$$

where $\frac{N'_k}{N_k}\epsilon_{bias,s^n_l}$ is the bias error introduced by replacing the interference term in recycled probability $p^k_R(s^n_l)$ with p_{unif}. And $\epsilon_{stat.,s^n_l}$ is the statistical error from estimating the recycled probabilities from a finite number of samples. These new expressions can then be solved to generate the mitigated outputs

$$p_{\text{miti}}(s^n_l) = \binom{m-n+k}{k}\left|\tilde{p}^k_R(s^n_l) - \frac{N'_k}{N_k} p_{unif}\right|. \tag{28}$$

We now provide pseudocode with the steps required to use the linear solving method to generate the mitigated output.

Algorithm 2: Linear solving method

input : Recycled probability estimator $\tilde{p}^k_R(s^n_l)$, and uniform probability p_{unif}.

1 Initialise variable $p_{\text{miti}}(s^n_l) \leftarrow 0$ for the mitigated output to be computed.

2 Update mitigated output variable as $p_{\text{miti}}(s^n_l) \leftarrow \binom{m-n+k}{k}\left|\tilde{p}^k_R(s^n_l) - \frac{N'_k}{N_k} p_{unif}\right|$.

output: Mitigated output $p_{mit}(s^n_l)$

There exists a regime of recycling mitigation usefulness, where the combined bias and the statistical errors present in the mitigated probabilities are lower than the statistical errors of the postselected distribution. Indeed, from Equation 28, it can be seen that the bias error of linear solving is upper bounded by

$$M_{bias} \in O\left(\binom{m-n+k}{k}\right)\epsilon_{bias,s^n_l},$$

we formally prove this in Appendix 12. Similarly, and as stated in Section 3, we show in Appendix 12 that statistical errors for a recycled probability constructed from $n-k$ photon statistics in linear solving is upper bounded by $M_{stat,k} \in O\left(\sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}\right)$, the $\binom{m}{n}$ can be replaced with N_s, if one is interested in mitigating a subset N_s of probabilities. With these upper bounds in hand, one can find a condition on the number of samples up to which recycling mitigation outperforms postselection by solving for N_{tot} in the following inequality

$$M_{bias} + M_{stat,k} \leq M_{stat,k=0}.$$

More details on this can be found in Appendix 12.

We now introduce the notion of dependency. This quantifies the correlation of the interference terms with the ideal probability within the recycled probabilities. A positive correlation means that the signal for the ideal probability is greater than $\binom{m-n+k}{k}^{-1}$, which can be used to improve the performance of linear solving. Each recycled probability can be rewritten to include a dependency term $d_k(s_l^n)$ quantifying this correlation, by reformulating the interference term as a linear function of $p(s_l^n)$ and p_{unif}. The expression

$$I_{s_l^n,k} = \big(1 - d_k(s_l^n)\big)p_{unif} + d_k(s_l^n)p(s_l^n) \tag{29}$$

defines the dependency term $d_k(s_l^n)$ of each recycled probability, and $k \leq n-1$. An average dependency term over the distribution, denoted d_k, may be calculated from the recycled probabilities (see eqn. 12.3.1 and appendix 12). Each recycled probability constructed from the $n-k$ output photon statistics may then be expressed in the form

$$\tilde{p}_R^k(s_l^n) = \frac{p(s_l^n)}{\binom{m-n+k}{k}} + \frac{N_k'}{N_k}\Big(\big((1-d_k)p_{unif} + d_k p(s_l^n) + \epsilon_{bias,s_l^n}\big) + \epsilon_{stat.,s_l^n}, \tag{30}$$

where ϵ_{bias,s_l^n} is the bias error introduced by replacing the interference term in recycled probability $p_R^k(s_l^n)$ with $\big((1-d_k)p_{unif} + d_k p(s_l^n)\big)$. Note that if the computed estimator for d_k is negative or greater than 1 then the dependency approach should be aborted and the original version of linear solving used. We conjecture it is always the case that $1 \geq d_k \geq 0$. The following pseudocode details the steps required to perform the linear solving with dependency method and generate the mitigated output.

We note that although the upper-bounds derived in Appendix 12 on the statistical and bias errors of linear solving with dependency are similar to those for linear solving without dependency, nevertheless numerical simulations show that linear solving with dependency reliably outperforms linear solving without dependency (see Fig. 3 (a) and (b)).

Algorithm 3: Linear solving with dependency method

input : Recycled probability estimator $\tilde{p}_R^k(s_l^n)$, uniform probability p_{unif}, an estimator for the absolute average deviation for the n output photon distribution $\tilde{D}_0$, and an estimator for the absolute average deviation for the $n-k$–photon recycled distribution $\tilde{D}_k$ (eqn. (32)).

1 Initialise variables for the mitigated output $p_{mit}(s_l^n) \leftarrow 0$ and the average dependency term $d_k \leftarrow 0$.

2 Update the average dependency term variable $d_k \leftarrow \frac{1}{\binom{m-n+k}{k}-1}\left(\frac{\binom{m-n+k}{k}\tilde{D}_k}{\tilde{D}_0} - \frac{1}{\binom{m-n+k}{k}}\right)$.

3 Update the mitigated output variable $p_{mit}(s_l^n) \leftarrow \left|\frac{\tilde{p}_R^k(s_l^n) + \frac{N_k'}{N_k}\left((-1+d_k)p_{unif}\right)}{\binom{m-n+k}{k}^{-1} + \frac{N_k'}{N_k}d_k}\right|$

output: Mitigated output $p_{mit}(s_l^n)$

4.4.2 Extrapolation

We now present methods by which extrapolation may be used to generate loss-mitigated outputs. From the definition of the recycled probability,

$$p_R^k(s_l^n) = \frac{p(s_l^n)}{\binom{m-n+k}{k}} + \frac{N'_k}{N_k} I_{s_l^n,k}, \tag{31}$$

the magnitude of the ideal probability signal within the recycled probabilities is proportional to $\binom{m-n+k}{k}^{-1}$. As k increases, the ideal signal magnitudes decrease as the recycled distributions converge towards uniform. The rate of decay of the ideal signal with k can be computed and then used to extrapolate mitigated outputs. We present two variations of extrapolation in which different types of dependence of the ideal probability signal on number of lost photons k are considered. A linear dependence is used for a linear extrapolation method, and an exponential dependence for an exponential extrapolation method. These two functions were considered primarily for heuristic reasons as seeming to reflect the decay behaviour observed in the recycled distributions. Both the linear and exponential extrapolation methods involve two iterations of optimisation. The first iteration computes an average decay parameter using the set of average absolute deviations of the recycled distributions. Where, for the set of recycled probabilities $\{p_R^k(s_l^n)\}_l$ constructed from k photon statistics, the average absolute deviation is defined as

$$D_k := \binom{m}{n}^{-1} \sum_l \left| p_R^k(s_l^n) - p_{unif} \right|. \tag{32}$$

The second iteration then uses the decay parameter to compute the mitigated values.

Linear extrapolation applies a linear model function to compute mitigated outputs. Here the least squares method is used to identify optimal parameters to fit a linear model function to data. Parameter optimisation is performed by minimising the sum of the squared residuals, where a residual is the difference between a data point and the model. A data set of N points is denoted $\{x_i, y_i\}_{i=1}^N$, where $\{x_i\}_{i=1}^N$ are the independent variables and $\{y_i\}_{i=1}^N$ are the dependent variables. The model function $f(x, \boldsymbol{\alpha})$ is optimised by varying the $\boldsymbol{\alpha}$ parameters to approximate the relation between independent and dependent variables found in the data set. The residual for each data point is defined $r_i := y_i - f(x_i, \boldsymbol{\alpha})$. The sum of the squared residuals is minimised to generate the optimal parameters

$$\boldsymbol{\alpha}_{\min} = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{N} r_i^2, \tag{33}$$

which are used to generate the optimised model function $f(x, \boldsymbol{\alpha}_{\min})$. This can then be used to make predictions about data outside the range of the data set used for optimisation.

In linear extrapolation, the linear model function used for the first iteration of linear least squares is

$$f(x, g_{\text{avg}}) = -g_{\text{avg}} x + \tilde{D}_0, \tag{34}$$

where $\tilde{D}_0$ is the average absolute deviation of the n-photon distribution from uniform computed from the statistics corresponding to postselecting on detecting all n photons, and g_{avg} is the optimal global linear decay parameter. The set $\{k, \tilde{D}_k\}_{k=1}^K$ is used as the data set to compute g_{avg}. Where $\{\tilde{D}_k\}_{k=1}^K$ is the set of average absolute deviations from uniform for the different distributions, for $K \leq n$, and $\tilde{D}_k$ is the estimated (from statistics where k photons were lost) absolute average deviation for the $n-k$–photon recycled distribution. After the optimal decay parameter is identified, another iteration of least squares is performed with updated model functions this time to generate the mitigated values. The data set used in this step is $\{k, \tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^K$. For the second iteration of linear least squares, each output bit string is assigned a linear model function of the form

$$f_{s_n}(x, \alpha_{s_l^n}) = \text{sgn}(p_{unif} - \tilde{p}_R^{k=1}(s_l^n)) g_{\text{avg}} x + \alpha_{s_l^n}. \tag{35}$$

For each output bit string s_n an optimal $\alpha_{s_l^n}$ is computed, generating the set $\{\alpha_{s_l^n}\}_l$, and the set of mitigated outputs is then $\{\alpha_{s_l^n} + p_{unif}\}_l$. In Appendix 13, error bounds are derived for the use of linear extrapolation to perform recycling mitigation, these indicate the existence of a regime where

the mitigation outperforms postselection. As with the analytical linear solving error bounds, the linear extrapolation error bounds allow the computation of an estimate of the number of samples up to which recycling mitigation with linear extrapolation outperforms postselection by solving for N_{tot} in the following inequality

$$M_{bias} + M_{stat,k} \leq M_{stat,k=0}.$$

We now provide pseudocode for applying the linear extrapolation method to generate mitigated outputs.

Algorithm 4: Linear extrapolation method

input : The number of data points $n_d \in \{n_d \in \mathbb{Z}^+ | n_d < n\}$ to be used in both iterations of least squares, the data set $\{k, \tilde{D}_k\}_{k=1}^{n_d}$ used to compute the gradient parameter g_{avg} in the first iteration of least squares, $\tilde{D}_0$, and the data set $\{k, \tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^{n_d}$ used to compute the mitigated output in the second iteration of least squares.

1 Initialise an average decay parameter variable $\tilde{g}_{\text{avg}} \leftarrow 0$, a prefactor variable $\alpha_{s_l^n} \leftarrow 0$, and a mitigated output variable $p_{mit}(s_l^n) \leftarrow 0$.

2 Use least squares method with model function $f(x_i, g_{\text{avg}}) = -\tilde{g}_{\text{avg}} x_i + \tilde{D}_0$ and data set $\{x_i, y_i\}_{i=1}^{n_d} := \{k, \tilde{D}_k\}_{k=1}^{n_d}$ to compute the value of the average decay parameter (slope), and assign this to variable $\tilde{g}_{\text{avg}}$.

3 Use least squares method with model function $f_{s_l^n}(x_i, \alpha_{s_l^n}) = \text{sgn}(p_{unif} - \tilde{p}_R^{k=1}(s_l^n))\tilde{g}_{\text{avg}} x_i + \alpha_{s_l^n}$ and data set $\{x_i, y_i\}_{i=1}^{n_d} := \{k, \tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^{n_d}$ to compute the value of the y-axis intercept, and assign this to variable $\alpha_{s_l^n}$.

4 Update mitigated output variable as $p_{mit}(s_l^n) \leftarrow p_{unif} + \alpha_{s_l^n}$.

output: Mitigated output $p_{mit}(s_l^n)$

The method for exponential extrapolation is broadly similar. However, non-linear least squares or a non-linear numerical optimisation method (e.g. the Levenberg–Marquardt algorithm [65, 66]) is instead used to compute the decay factor and the mitigated probabilities. The exponential model function used for the first step is

$$f(x, \alpha_{\text{avg}}) = \tilde{D}_0 e^{-\alpha_{\text{avg}} x}, \tag{36}$$

where α_{avg} is the optimal global exponential decay parameter [3]. Where again the set $\{k, \tilde{D}_k\}_{k=1}^{K}$ is used as the data, this time to compute the optimal value of α_{avg}. And then each output bit string for which a mitigated probability is to be generated is assigned a model function of the form

$$f_{s_n}(x, \Lambda_{s_n}) = \Lambda_{s_l^n} e^{-\alpha_{\text{avg}} x} + p_{unif}. \tag{37}$$

An optimal prefactor, denoted $\Lambda_{s_l^n}^{\text{opt}}$, is computed for each bit string s_l^n using numerical optimisation. This generates the set of prefactors $\{\Lambda_{s_l^n}^{\text{opt}}\}_l$, and the set of mitigated outputs is then $\{\Lambda_{s_l^n}^{\text{opt}} + p_{unif}\}_l$. The following pseudocode gives the steps required to use the exponential extrapolation method to generate the mitigated outputs.

[3] Note that, in order to satisfy $f(n, \alpha_{\text{avg}}) = 0$, the fitting function should be $f(x, \alpha_{\text{avg}}) = \tilde{D}_0 e^{-\alpha_{\text{avg}} x} + \beta$ with $\beta < 0$. However, as we are intersted in the decay of $f(.)$ with increasing k, and we look at (far from n) values of $k \in O(1)$ in the fitting, we can exclude β in the above expression.

Algorithm 5: Exponential extrapolation method

input : The number of data points $n_d \in \{n_d \in \mathbb{Z}^+ | n_d < n\}$ to be used in both iterations of least squares, the set $\{k, \tilde{D}_k\}_{k=1}^{n_d}$ used to compute the gradient parameter g_{avg} in the first iteration of least squares, $\tilde{D}_0$, and the data set $\{k, \tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^{n_d}$ used to compute the prefactor value in the second iteration of least squares.

1 Initialise an average decay parameter variable $\alpha_{avg} \leftarrow 0$, a prefactor variable $\Lambda_{s_l^n} \leftarrow 0$, and a mitigated output variable $p_{mit}(s_l^n) \leftarrow 0$.

2 Use least squares method with model function $f(x, \alpha_{\text{avg}}) = \tilde{D}_0 e^{-\alpha_{\text{avg}} x}$ and data set $\{x_i, y_i\}_{i=1}^{n_d} := \{k, \tilde{D}_k\}_{k=1}^{n_d}$ to compute the value of the average decay parameter and assign this to variable α_{avg}.

3 Use least squares method with model function $f_{s_n}(x, \Lambda_{s_l^n}) = \Lambda_{s_l^n} e^{-\alpha_{\text{avg}} x} + p_{unif}$ and data set $\{x_i, y_i\}_{i=1}^{n_d} := \{k, \tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^{n_d}$ to compute the value of the prefactor and assign this to variable $\Lambda_{s_l^n}$.

4 Update mitigated output variable as $p_{mit}(s_l^n) \leftarrow p_{unif} + \Lambda_{s_l^n}$.

output: Mitigated output $p_{mit}(s_l^n)$

Note that the ideal signal magnitude decays proportionally with $\binom{m-n+k}{k}^{-1} \sim m^{-k}$ in eqn. 31, which intuitively motivates the choice of an exponential model function to reflect this decay behaviour. In the next sections, we provide numerical evidence indicating that there exists a non-trivial sampling regime where exponential extrapolation outperforms postselection. We also note that in the numerical simulations, extrapolation using an exponential model function consistently outperforms linear extrapolation (see Fig. 3 (c) and (d)). This may be a consequence of the exponential model function better reflecting the ideal signal decay behaviour.

As a final remark we comment on how one can estimate D_k in practice. In [67], it was shown that D_0 is lower bounded by a constant. The results of numerical simulations plotted in Fig. 2 highlight that the same appears to hold for D_k when $k \in O(1)$. Intuitively, this makes sense as D_k, similar to D_0, are expected to contain probabilities that are hard to simulate classically (see Section 6) , and are thus far from the uniform distribution. In practice, this means that in order to estimate D_k to a good ($\epsilon \in O(1)$) precision, one only needs to compute an $O(1)$ fraction of probabilities $p_R(s_l^n)$, with s_l^n picked uniformly randomly, and compute the mean of $|p_R(s_l^n) - p_{unif}|$. This gives a high-confidence estimate of D_k by Hoeffding's inequality. Consequently, this means that estimating the D_k to be used in extrapolation recycling mitigation does not require significantly more statistics than what is needed in linear solving.

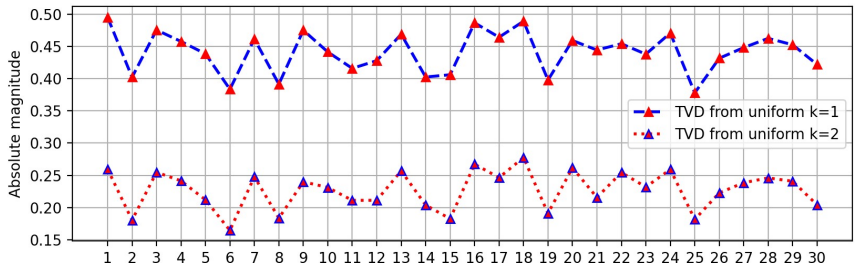
4.5 Normalising the mitigated distribution

The postprocessing techniques we have presented take as input a set of recycled probabilities $\{p_R(s_l^n)\}_l$ and output a set of mitigated values $\{p_{mit}(s_l^n)\}_l$, or a subset of these values. We have made the distinction between *values* and *probabilities*, as the outputs $\{p_{mit}(s_l^n)\}_l$ satisfy $p_{mit}(s_l^n) > 0$ but are not normalised in general. That is, $\sum_l p_{mit}(s_l^n) = N$ with no guarantee that $N = 1$. Analytical and numerical calculations performed in previous sections have shown that, with high probability,

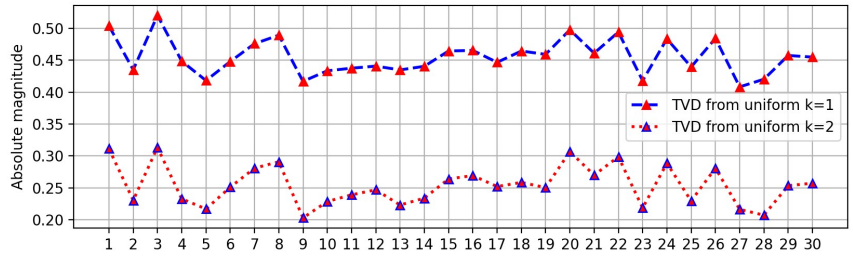
$$||\vec{p}_{mit} - \vec{p}_{id}||_1 \leq ||\vec{p}_{post} - \vec{p}_{id}||_1,$$

and $|p_{mit}(s_l^n) - p(s_l^n)| \leq |p_{post}(s_l^n) - p(s_l^n)|$ up to a certain number of samples. $\vec{p}_{mit}$, $\vec{p}_{id}$, and $\vec{p}_{post}$, are vectors containing respectively the mitigated values, the ideal n–photon probabilities, and the probabilities obtained by postselection, and $||.||_1$ is the usual l_1 norm, $||\vec{v}||_1 := \sum_i |v_i|$.

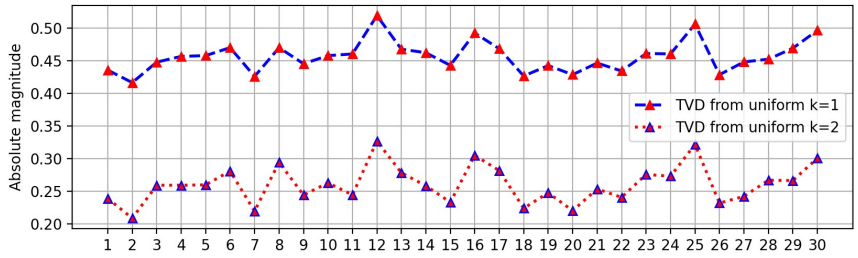
In practice, one is usually interested in computing the expectation value of some observable O, defined as $\langle \mathsf{O} \rangle = \sum_i p_i w_i$, where w_i are some weights, and $\{p_i\}$ a subset of probabilities of the quantum circuit. In this case we can replace the p_i's by the unnormalised mitigated values and obtain guarantees similar to those stated above. However, if we are interested in mitigating the entire distribution, then we would need to normalise $\vec{p}_{mit}$. The easiest way to do this, and what we do for our numerical simulations, is to define, $\vec{p}_{mit,nor} := \frac{1}{N} \vec{p}_{mit}$. One can easily check that $||\vec{p}_{mit,nor}||_1 = 1$, and therefore that it is a vector of normalised probabilities.



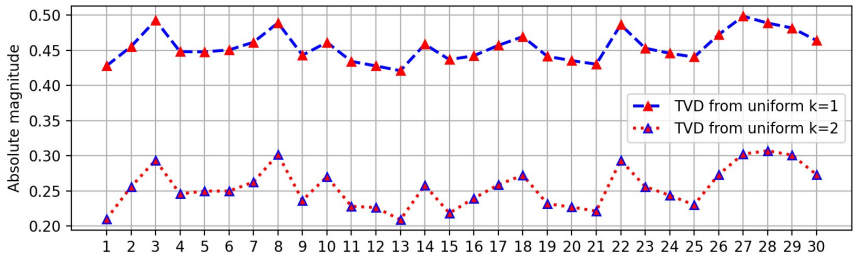
(a)



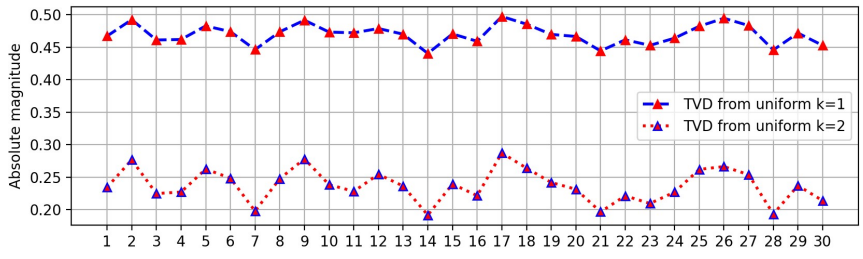
(b)



(c)



(d)



(e)

Figure 2: The results of numerical simulations where D_k for $k \in \{1, 2\}$ is plotted for 30 random unitaries with $m = 20$ for (a) $n = 3$, (b) $n = 4$, (c) $n = 5$, (d) $n = 6$, and (e) $n = 7$. From these results, as with [67] where it was shown that D_0 is lower bounded by a constant, it appears this also holds for D_k when $k \in O(1)$.

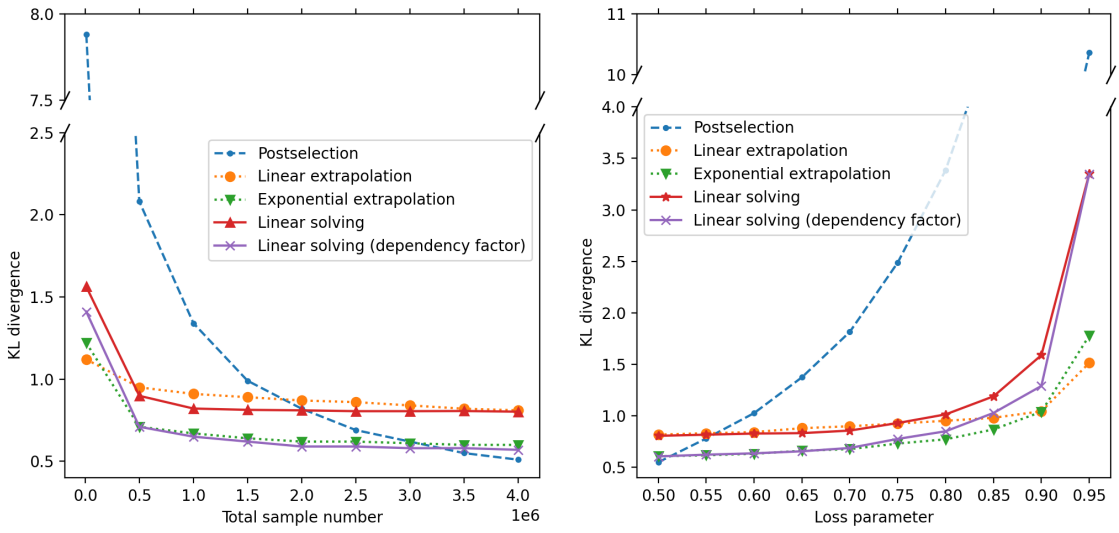


Figure 3: A numerical performance comparison of the different methods of recycling mitigation and postselection for random unitary circuits with $m = 20$ modes and $n = 4$ photons. (a) For a uniform loss parameter of $\eta = 0.8$, the KL divergence from the ideal output distribution for linear extrapolation, exponential extrapolation, linear solving and linear solving with dependency versions of recycling mitigation and postselection is plotted against total sample number. (b) For a total number of samples of $N_{tot} = 1 \times 10^5$ and a uniform loss parameter in the range $\eta \in [0.5, 0.9]$, the KL divergence from the ideal output distribution of linear extrapolation, exponential extrapolation, linear solving and linear solving with dependency versions of recycling mitigation and postselection is plotted against total sample number.

We will show that such a normalisation provably produces a distribution $\vec{p}_{mit,nor}$ which is closest possible to $\vec{p}_{mit}$ in l_1-norm.

Let $\vec{q}_m$ be a vector of positive values which is a *closest* (in l_1-norm) normalised vector to $\vec{p}_{mit}$. More precisely, $\vec{q}_m$ satisfies

$$||\vec{q}_m - \vec{p}_{mit}||_1 = \min_{\vec{q}\,|\,||\vec{q}||_1=1} ||\vec{q} - \vec{p}_{mit}||_1.$$

An immediate observation, by definition, is that

$$||\vec{q}_m - \vec{p}_{mit}||_1 \leq ||\vec{p}_{id} - \vec{p}_{mit}||_1.$$

We now prove

Lemma 15. *A valid choice of $\vec{q}_m$ is $\vec{q}_m = \vec{p}_{mit,nor}$.*

Proof. $||\vec{p}_{mit,nor} - \vec{p}_{mit}||_1 = ||\vec{p}_{mit,nor} - \vec{p}_{mit,nor}N||_1 = |1 - N| = \left|1 - ||\vec{p}_{mit}||_1\right|$.

For any normalized vector of positive values $\vec{q}$, we can use a reverse triangle inequality to show $||\vec{q} - \vec{p}_{mit}||_1 \geq \left|||\vec{q}||_1 - ||\vec{p}_{mit}||_1\right| \geq \left|1 - ||\vec{p}_{mit}||_1\right| \geq ||\vec{p}_{mit,nor} - \vec{p}_{mit}||_1$. Thus $\vec{p}_{mit,nor}$ is a valid choice of $\vec{q}_m$, by definition of $\vec{q}_m$. □

5 Numerical simulations

In the numerical simulations recycling mitigation is used to mitigate the effects of uniform photon loss error on the output of an otherwise ideal simulation of a linear optical quantum circuit. In the simulations, random matrices are chosen for each experiment which are decomposed and implemented with a linear interferometer, and a uniform photon loss model is applied with a defined loss parameter. The simulations were performed using Perceval [51], a pythonic framework for the simulation of photonic quantum circuits. Uniform photon loss channels commute with the interferometer, and so all loss channels, including loss due to imperfect sources and measurement,

can be propagated to the end of the circuit. And so the effects of loss may be modelled by an ideal photon source, an ideal interferometer M, and a combined photon loss channel acting immediately before an ideal measurement operation. This means that rather than an output photon from the circuit incident on a detector being in the state $|1\rangle\langle 1|$, it is instead

$$|1\rangle\langle 1| \to (1-p)\,|1\rangle\langle 1| + p\,|0\rangle\langle 0|\,.$$

With the probability of photon loss, p, the same for all output modes. Noise of this form may be considered analogous to the types of measurement noise commonly considered in circuit model quantum computing, as this is often modelled as an error channel followed by an ideal measurement.

For all the experiments in Fig. 3, parameter settings of 20 modes and 4 photons were used. Fig. 3 (a) plots the performance of linear solving recycling mitigation, both with and without using the dependency factor, and postselection against the total number of samples used for a fixed uniform loss parameter of $\eta = 0.8$. Linear solving outperformed postselection up to $\approx 1.1 \times 10^6$ samples, while linear solving with dependency outperformed postselection up to $\approx 3.1 \times 10^6$ samples. Note that $\binom{20}{4}^2 \approx 23 \times 10^6$, and thus recycling mitigation is outperforming postselection for sample sizes of the order of $\binom{m}{n}^2 = \binom{20}{4}^2$. Fig. 3 (b) plots the performance of the linear solving methods and postselection against changing loss parameter for a fixed number of samples of 1×10^5. Linear solving outperformed postselection for loss above ≈ 0.6, and linear solving with the dependency factor above loss of ≈ 0.5. Fig. 3 (c) plots the performance of linear extrapolation, exponential extrapolation and postselection against the total number of samples used for a fixed uniform loss parameter of $\eta = 0.8$. Linear extrapolation outperformed postselection up to $\approx 1.7 \times 10^6$ samples, while exponential extrapolation does so up to $\approx 3.0 \times 10^6$ samples. Fig. 3 (d) plots the performance of the extrapolation methods and postselection against changing loss parameter for a fixed number of samples of 1×10^5. Linear extrapolation outperformed postselection for loss above ≈ 0.5, and linear solving with the dependency factor above loss of ≈ 0.5. These results indicate the exponential function is a better model for the physical behaviour of the decay than the linear function, we believe this may be explained in terms of the prefactor of the ideal output probability in eqn. 31 decreasing exponentially (specifically $\propto (m-n+k)^{-k}$) with increasing k. In current quantum linear optical devices losses including source, interferometer and measurement are commonly above 50% [49]. Therefore, the evidence from these simulations indicates that recycling mitigation may be usefully applied to the current generation of linear optical devices. Also, in the previous section an upper bound on the number of samples was computed for a particular experiment.

An intuitive explanation for the results in Fig. 3 is that the mitigated outputs have a lower statistical error and so converge more quickly than the postselected outputs for increasing sample number and decreasing loss parameter. Furthermore, the bias errors in the mitigated outputs mean that they do not converge to the ideal outputs, whereas the postselected outputs are unbiased estimators. And so for sufficiently high sample number or low loss parameter postselection will eventually outperform recycling mitigation. There are, however, large sampling and loss regimes for which recycling mitigation reliably outperforms postselection.

Theorems 8 and 10 in section 4.3 provide statistical upper bounds on the deviation of the interference terms from p_{unif}. An inspection of these bounds shows that this deviation is with high confidence $\Omega(\sqrt{p_{unif}})$.

We conjecture that this bound could be considerably tightened. In support of this we ran numerical experiments for computations of size $m = 16$ modes and $n = 4$ photons, computing the average of $|I_{s_l^n,k} - p_{unif}|$ over all bit strings $\{s_l^n\}_l$. This computation was repeated for 20 randomly selected linear optical intereferometers U. This data is plotted in Fig. 4. As can be seen from the figure it seems that the analytically computed bounds on the bias error are loose. Indeed, it seems that $|I_{s_l^n,k} - p_{unif}| \in o(p_{unif})$, in line with conjecture 3.

Finally, note that $|I_{s_l^n,k} - p_{unif}|$ seems to be generally decreasing with increasing k. However, $\binom{m-n+k}{k}|I_{s_l^n,k} - p_{unif}|$ actually increases with increasing k, as can be verified by a direct calculation using data from Figure 4. Ultimately, since $\binom{m-n+k}{k}|I_{s_l^n,k} - p_{unif}|$ is directly related to the bias error incurred when computing the mitigated probabilities from the recycled probabilities (see appendix 12), increasing values of k in the recycling mitigation will lead to worse performances, as predicted by Lemma 6.

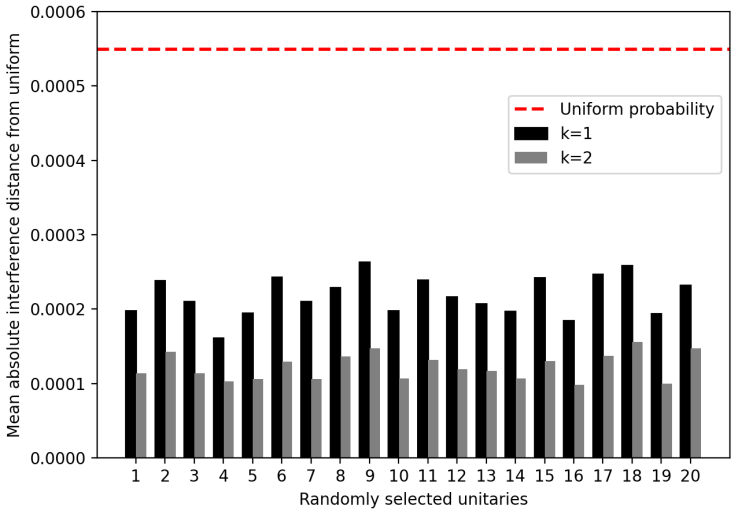


Figure 4: The mean absolute value of the distance of the interference terms from the uniform probability were computed for 20 randomly selected unitaries. For all unitaries, and for $k = 1$ and $k = 2$, the magnitude of the computed values were observed to be exponentially small in terms of m and n (being of the order of $\binom{m}{n}^{-1}$). This indicates that it may be possible to derive tighter analytical bounds than those stated in thm. 8 and thm. 10.

6 Properties of the mitigated probabilities in the absence of statistical error

As seen previously, recycling mitigation is a biased error mitigation. Given a U and a bit string s, this means that in the limit of infinite samples the mitigated probability $p(s)$ is related to the ideal probability $p_{id}(s)$ by

$$|p(s) - p_{id}(s)| = \epsilon_{mitbias,s}.$$

Numerical simulations in Fig. 5 using the exponential extrapolation recycling mitigation technique with values of k independent of m, n show that $E_s(\epsilon_{mitbias,s}) < p_{unif}$ both in the case of fixed m and varying n, as well as fixed n and varying m. Although the plots in Fig. 5 are for a fixed U, we observe similar curves when testing with different values of U. Our numerical simulations suggest that Conjecture 3 is true. We will now assume that this conjecture is true and derive from it a condition on the mitigated probabilities.

Let $m = \Omega(n^5)$, then $p_{id}(s) \approx \frac{|\mathsf{Per}(X_s)|^2}{m^n}$ with X_s an i.i.d Gaussian matrix [6], also $p_{unif} \approx \frac{n!}{m^n}$. Thus, Conjecture 3 implies, by using a Markov inequality with $\alpha > 1$,

$$Pr_s(|m^n p(s) - |\mathsf{Per}(X_s)|^2| \leq \delta(n)\alpha n!) \geq 1 - \frac{1}{\alpha}.$$

Choosing α such that $\alpha\delta(n) := \gamma(n) < 1$, we have that

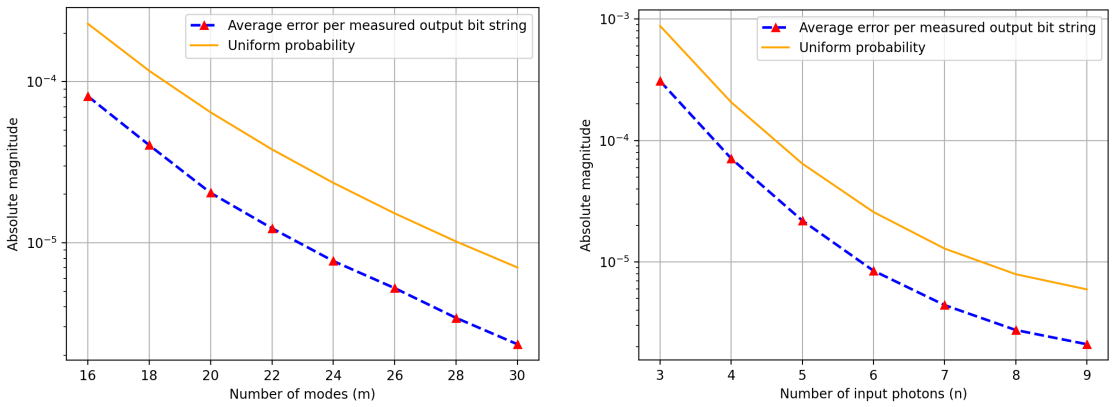
$$Pr_s(|m^n p(s) - |\mathsf{Per}(X_s)|^2| \leq \gamma(n) n!) \geq 1 - \frac{1}{\alpha}.$$

Since this equation holds with high probability p over the choice of U, then we can say that

$$Pr_{X \in \mathcal{G}^{n\times n}}(|m^n p(s) - |\mathsf{Per}(X)|^2| \leq \gamma(n) n!) \geq (1 - \frac{1}{\alpha})p, \tag{38}$$

where $\mathcal{G}^{n\times n}$ is the ensemble of $n \times n$ Gaussian matrices.

We will now argue that the existence of a $\mathsf{poly}(n)$-time classical algorithm C that can compute any mitigated probability is an unlikely complexity theoretic conequence. Indeed, if C existed, then eqn. 38 directly implies that C can be used to solve the $|GPE_{\pm}|^2$ problem with some restrictions on the parameters [6]. More precisely, computing the mitigated probabilities allows us to estimate $|\mathsf{Per}(X)|^2$ to additive error $\gamma(n)n! < n!$ with probability $(1 - \frac{1}{\alpha})p \approx 1 - \frac{1}{\alpha}$ over $X \in \mathcal{G}^{n\times n}$. The



(a) (b)

Figure 5: The numerical simulations that produced these plots involved exact computation of the mitigated probabilities and so only the bias error is present. The magnitude of the average error per output bit string is compared with that of the uniform probability. For all chosen parameters the average error per bit string is below the uniform probability. (a) The average error per bit string plotted for a range of modes $[16, 30]$ for a fixed number of input photons $n = 5$. (b) The average error per bit string plotted for a range of input photons $[3, 9]$ for a fixed number of modes photons $m = 20$.

$|GPE_{\pm}|^2$ is about classically estimating $|\mathsf{Per}(X)^2|$ to within additive error $\epsilon n!$ and probability $1-\delta$ over $X \in \mathcal{G}^{n\times n}$ for any ϵ, δ, and is strongly believed to be $\sharp$P-hard. Furthermore, in [58] it was argued that $|GPE_{\pm}|^2$ with some restrictions on the parameters similar to those in eqn. 38 should also be hard to do classically. This means that there likely is no such classical algorithm C.

In conclusion, despite recycling mitigation being a biased error mitigation technique, we have presented numerical evidence to argue that the bias errors are small enough to render the mitigated probabilities hard to compute classically.

7 Evidence that zero noise extrapolation (ZNE) techniques present no advantage over postselection

In this section, we provide strong evidence that techniques based on ZNE when applied to mitigating photon loss in DVLOQC in general provide no advantage over postselection.

Suppose we are interested in computing a specific marginal probability $p(n_1...n_l|n)$ of observing n_i photons in mode $i \in \{1, \ldots, l\}$, with $l \leq m$, and $\sum_{i=1,\ldots,l} n_i = c$, with $c \leq n$. The notation $|n$ indicates that we are computing the *ideal* marginal probability, when no photon is lost. Let $p(n_1 \ldots n_l \cap j)$ be the probability of observing the output $(n_1, \ldots, n_l)$ and detecting j photons in all m modes, with $j \in \{c, \ldots, n\}$. When postselecting on no photons being lost, we are computing

$$p(n_1 \ldots n_l \cap n) = (1-\eta)^n p(n_1 \ldots n_l|n).$$

However, if we compute $p(n_1 \ldots n_l)$ without caring about whether no photon is lost, we end up computing

$$p_\eta(n_1...n_l) = \sum_{i=0,\ldots,n-c} (1-\eta)^{n-i}\eta^i p(n_1 \ldots n_l|n-i). \tag{39}$$

Extrapolation techniques consist of estimating $p_\eta(n_1 \ldots n_l)$ for different values $\{\eta_i\}$ of loss, then deducing from these an estimate of $p(n_1 \ldots n_l|n)$. One example of how this can be done is the Richardson extrapolation technique, at the heart of the zero noise extrapolation (ZNE) approach [14]. Interestingly, it is possible to derive a condition indicating that these techniques offer no advantage over postselection in terms of estimating $p(n_1 \ldots n_l|n)$. Let $\eta = \mathsf{min}_i \eta_i$, and suppose we collect $O(\frac{1}{\epsilon^2_{max}})$ samples from our device, where $0 < \epsilon_{max} \leq 1$, then $O\big(\frac{(1-\eta)^n}{\epsilon^2_{max}}\big)$ of these will be samples where no loss has occurred. Using these postselected

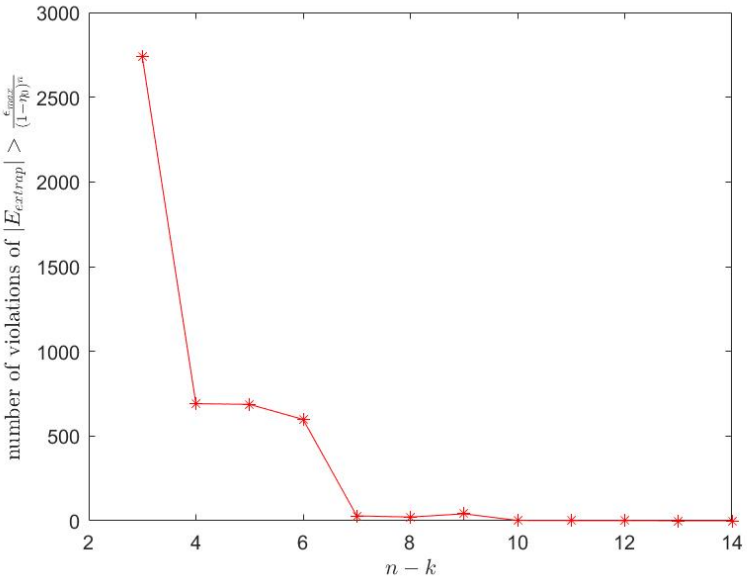


Figure 6: Number of violations of Thm. 16 inequality plotted versus $n-c$.

samples one can compute, with high confidence, an estimate $\tilde{p}(n_1 \dots n_l|n)$ of $p(n_1 \dots n_l|n)$ such that $|\tilde{p}(n_1 \dots n_l|n) - p(n_1 \dots n_l|n)| \leq \frac{\epsilon_{max}}{\sqrt{(1-\eta)^n}}$, by Hoeffding's inequality. For Richardson extrapolation, in Appendix 15 we use $O\big(n\frac{1}{\epsilon_{max}^2}\big)$ samples, and compute an estimate $\tilde{p}_{extrap}(n_1 \dots n_l|n)$, with $E_{extrap} := |\tilde{p}_{extrap}(n_1 \dots n_l|n) - p(n_1 \dots n_l|n)|$ the error incurred. We then explicitly compute $\mathsf{M}(E_{extrap})$, an upper bound on E_{extrap}, in terms of ϵ_{max} and the η_i's, and show the following

Theorem 16. *For all $n \geq n_0$, with n_0 a positive integer,* $\mathsf{M}(E_{extrap}) \geq \frac{\epsilon_{max}}{\sqrt{(1-\eta)^n}}$.

Thm. 16, proven in Appendix 15, is strong evidence that techniques based on Richardson extrapolation offer no advantage over postselection. Our main technical contribution in proving Thm. 16 is to link determining the error of ZNE methods to computing the norm of the inverse of a Vandermonde matrix whose entries are determined by η, we then use existing results to upper bound this norm [39].

To strengthen our theoretical analysis, we also numerically compute the number of violations of the inequality $E_{extrap} \geq \frac{\epsilon_{max}}{\sqrt{(1-\eta_0)^n}}$ and plot these in Figure 6. We took $\epsilon_{max} = 0.01$, $\eta_0 = 0.01$, $\eta_{n-c} = 0.95$, and η_i for $1 < i < n-c$ equally spaced. We varied the value of $n-c$ between 3 and 14, with $c = \mathsf{ceil}(n/3)$ and $\mathsf{ceil}(.)$ is the ceiling function. For each value of $n-c$ we performed 3000 runs, where at each run we took $n-c+1$ values of $\{\epsilon_i\}$ chosen uniformly randomly from $[-\epsilon_{max}, \epsilon_{max}]$. As can be observed in Figure 6, the number of violations approaches zero with increasing n, confirming that extrapolation performs worse than post-selection after some value of n. A similar behaviour is observed for different values of ϵ_{max}, η_0 and η_{n-c}.

8 Discussion

In summary, we have presented a family of techniques, collectively referred to as *recycling mitigation*, for mitigating the effects of photon loss on the outputs of linear optical quantum circuits in the discrete variable setting. We provided analytical and numerical evidence that these techniques outperform postselection - currently the standard method for mitigating loss in linear optical circuits.

There are many possible directions of future work. For instance, carrying out of a rigorous analysis of the set-up where recycling mitigation is applied to linear optical computation with adaptive measurements and feedforward, such as in [68]. This would be interesting for a number of reasons, not least of which is that in this computational set-up, with a sufficient number of adaptive measurement steps, a quantum advantage for probability estimation is feasible [68]. As an estimate, if n_{ad} is the number of photons consumed in the r feed-forward steps in the absence of loss, and η is the loss per mode, the overall probability that the r feed-forward steps are executed correctly is $(1-\eta)^{n_{ad}}$. This introduces an unavoidable additional exponential sampling overhead to recycling mitigation. Nevertheless, it would be interesting benchmark the performance of recycling mitigation against only postselection using a set-up of this kind, especially in the possible regime where there is not sufficient feed-forward for universality, but where classical simulation methods are also inefficient [68].

It would also be interesting to investigate how the method might be combined with early fault-tolerance schemes [69] for optical quantum computing. Specifically, in investigating the resource trade-off between error mitigation and correction needed to achieve a certain quality of computational output.

Another interesting question, motivated by finding ways to eliminate the bias error in our introduced mitigation techniques is whether one can *exactly* represent an n–photon probability as a sum of $n-k$–photon probabilities. This can be done for the case of minors of unitary matrices with real and positive entries, such as those used to solve graph problems in DVLOQC [53]. As an example, consider a Laplace expansion of the permanent of an $n \times n$ minor $U_{\mathbf{s},\mathbf{t}} := (u_{ij})_{i,j\in\{1,\ldots,n\}}$ of a linear optical unitary U, this reads

$$\mathsf{Per}(U_{\mathbf{s},\mathbf{t}}) = u_{11}\mathsf{Per}\begin{pmatrix} u_{22} & \ldots & u_{2n} \\ . & . & . \\ . & . & . \\ . & . & . \\ u_{n2} & \ldots & u_{nn} \end{pmatrix} + \cdots + u_{1n}\mathsf{Per}\begin{pmatrix} u_{21} & \ldots & u_{2,n-1} \\ . & . & . \\ . & . & . \\ . & . & . \\ u_{n1} & \ldots & u_{n,n-1} \end{pmatrix}.$$

If $U_{\mathbf{s},\mathbf{t}}$ is composed of positive real entries, then $\mathsf{Per}(U_{\mathbf{s},\mathbf{t}})$ is proportional to the square root of a probability $\sqrt{p(s|t)}$ of a certain output of linear optical circuit U with n input photons [6]. In turn, each of the permanents of $n-1 \times n-1$ matrices on the right hand side of the above equation is proportional to the square root of a probability of a certain output of a linear optical circuit with $n-1$ input photons. Thus, in the case of minors with positive real entries, there is an exact way to represent an n–photon probability as a sum of $n-1$ photon probabilities. In this case as well, the process can be generalised, by successive Laplace expansions, to expressing n–photon probabilities exactly as a sum of $n-k$–photon probabilities.

This idea of exactly representing n–photon probabilities as sums of $n-k$–photon probabilities is interesting since, in the presence of photon loss, $n-k$–photon experiments in general produce more converged statistics than n–photon experiments, for a comparable number of runs of the experiments in both cases. Furthermore, since the n–photon probabilities are exactly expressible as a sum of $n-k$–photon probabilities, one can use the $n-k$–photon experiments to compute the n–photon probabilities and *indefinitely* (because of lower statistical error) outperform postselection. It is an interesting question to determine whether similar results hold for broader classes of matrices.

Finally, it could be interesting to extend our techniques to the case where we allow collisions in the output. The construction of the recycled probabilities presented generalises straightforwardly to the collision regime, the main technical challenge is therefore to adapt our proofs to this regime.

Code availability

Open-source code to perform recycling mitigation may be found at the following Github location:
https://github.com/Quandela/Perceval/tree/main/perceval/error_mitigation

Acknowledgments

The authors would like to thank Hugo Thomas for suggesting the approach taken to proving Lemma 15. Furthermore, the authors would like to thank Enguerrand Monard, Pierre-Emmanuel Emeriau, Shane Mansfield, Stephen Wein, Alexia Salavrakos, Emilio Annoni, Robert Booth, and Elham Kashefi for helpful discussions. The authors would also like to thank Raúl García-Patrón for providing helpful feedback on an earlier version of this manuscript. This work has been co-funded by the European Commission as part of the EIC accelerator program under the grant agreement 190188855 for SEPOQC project, by the Horizon-CL4 program under the grant agreement 101135288 for EPIQUE project, and by the OECQ Project financed by the French State as part of France 2030.

References

[1] E. Knill, R. Laflamme, and G. J. Milburn. "A scheme for efficient quantum computation with linear optics". Nature **409**, 46–52 (2001).

[2] R. Raussendorf, J. Harrington, and K. Goyal. "Topological fault-tolerance in cluster state quantum computation". New Journal of Physics **9**, 199 (2007).

[3] Grégoire de Gliniasty, Paul Hilaire, Pierre-Emmanuel Emeriau, Stephen C. Wein, Alexia Salavrakos, and Shane Mansfield. "A Spin-Optical Quantum Computing Architecture". Quantum **8**, 1423 (2024).

[4] Sara Bartolucci, Patrick Birchall, Hector Bombín, Hugo Cable, Chris Dawson, Mercedes Gimeno-Segovia, Eric Johnston, Konrad Kieling, Naomi Nickerson, Mihir Pant, Fernando Pastawski, Terry Rudolph, and Chris Sparrow. "Fusion-based quantum computation". Nature Communications **14**, 912 (2023).

[5] Robert Raussendorf and Hans J. Briegel. "A One-Way Quantum Computer". Physical Review Letters **86**, 5188–5191 (2001).

[6] Scott Aaronson and Alex Arkhipov. "The computational complexity of linear optics". In Proceedings of the forty-third annual ACM symposium on Theory of computing. Pages 333–342. STOC '11New York, NY, USA (2011). Association for Computing Machinery.

[7] Pascale Senellart, Glenn Solomon, and Andrew White. "High-performance semiconductor quantum-dot single-photon sources". Nature Nanotechnology **12**, 1026–1039 (2017).

[8] Michael Reck, Anton Zeilinger, Herbert J. Bernstein, and Philip Bertani. "Experimental realization of any discrete unitary operator". Physical Review Letters **73**, 58–61 (1994).

[9] Robert H. Hadfield. "Single-photon detectors for optical quantum information applications". Nature Photonics **3**, 696–705 (2009).

[10] Hui Wang, Wei Li, Xiao Jiang, Y.-M. He, Y.-H. Li, X. Ding, M.-C. Chen, J. Qin, C.-Z. Peng, C. Schneider, M. Kamp, W.-J. Zhang, H. Li, L.-X. You, Z. Wang, J.P. Dowling, S. Höfling, Chao-Yang Lu, and Jian-Wei Pan. "Toward Scalable Boson Sampling with Photon Loss". Physical Review Letters **120**, 230502 (2018).

[11] Daniel J. Brod, Ernesto F. Galvão, Andrea Crespi, Roberto Osellame, Nicolò Spagnolo, and Fabio Sciarrino. "Photonic implementation of boson sampling: a review". Advanced Photonics **1**, 034001 (2019).

[12] Jelmer Renema, Valery Shchesnovich, and Raul Garcia-Patron. "Classical simulability of noisy boson sampling" (2019). arXiv:1809.01953.

[13] Kristan Temme, Sergey Bravyi, and Jay M. Gambetta. "Error Mitigation for Short-Depth Quantum Circuits". Physical Review Letters **119**, 180509 (2017).

[14] Suguru Endo, Simon C. Benjamin, and Ying Li. "Practical Quantum Error Mitigation for Near-Future Applications". Physical Review X **8**, 031027 (2018).

[15] Tudor Giurgica-Tiron, Yousef Hindy, Ryan LaRose, Andrea Mari, and William J. Zeng. "Digital zero noise extrapolation for quantum error mitigation". In 2020 IEEE International Conference on Quantum Computing and Engineering (QCE). Pages 306–316. (2020).

[16] Andre He, Benjamin Nachman, Wibe A. de Jong, and Christian W. Bauer. "Zero-noise extrapolation for quantum-gate error mitigation with identity insertions". Physical Review A **102**, 012426 (2020).

[17] Armands Strikis, Dayue Qin, Yanzhu Chen, Simon C. Benjamin, and Ying Li. "Learning-Based Quantum Error Mitigation". PRX Quantum **2**, 040330 (2021).

[18] Andrea Mari, Nathan Shammah, and William J. Zeng. "Extending quantum probabilistic error cancellation by noise scaling". Physical Review A **104**, 052607 (2021).

[19] Ewout van den Berg, Zlatko K. Minev, Abhinav Kandala, and Kristan Temme. "Probabilistic error cancellation with sparse Pauli–Lindblad models on noisy quantum processors". Nature Physics **19**, 1116–1121 (2023).

[20] X. Bonet-Monroig, R. Sagastizabal, M. Singh, and T. E. O'Brien. "Low-cost error mitigation by symmetry verification". Physical Review A **98**, 062339 (2018).

[21] R. Sagastizabal, X. Bonet-Monroig, M. Singh, M. A. Rol, C. C. Bultink, X. Fu, C. H. Price, V. P. Ostroukh, N. Muthusubramanian, A. Bruno, M. Beekman, N. Haider, T. E. O'Brien, and L. DiCarlo. "Experimental error mitigation via symmetry verification in a variational quantum eigensolver". Physical Review A **100**, 010302 (2019).

[22] Thomas E. O'Brien, Stefano Polla, Nicholas C. Rubin, William J. Huggins, Sam McArdle, Sergio Boixo, Jarrod R. McClean, and Ryan Babbush. "Error Mitigation via Verified Phase Estimation". PRX Quantum **2**, 020317 (2021).

[23] Rawad Mezher, James Mills, and Elham Kashefi. "Mitigating errors by quantum verification and postselection". Physical Review A **105**, 052608 (2022).

[24] Bálint Koczor. "Exponential Error Suppression for Near-Term Quantum Devices". Physical Review X **11**, 031057 (2021).

[25] Bálint Koczor. "The dominant eigenvector of a noisy quantum state". New Journal of Physics **23**, 123047 (2021).

[26] William J. Huggins, Sam McArdle, Thomas E. O'Brien, Joonho Lee, Nicholas C. Rubin, Sergio Boixo, K. Birgitta Whaley, Ryan Babbush, and Jarrod R. McClean. "Virtual Distillation for Quantum Error Mitigation". Physical Review X **11**, 041036 (2021).

[27] Kaoru Yamamoto, Suguru Endo, Hideaki Hakoshima, Yuichiro Matsuzaki, and Yuuki Tokunaga. "Error-Mitigated Quantum Metrology via Virtual Purification". Physical Review Letters **129**, 250503 (2022).

[28] Jarrod R. McClean, Zhang Jiang, Nicholas C. Rubin, Ryan Babbush, and Hartmut Neven. "Decoding quantum errors with subspace expansions". Nature Communications **11**, 636 (2020).

[29] Nobuyuki Yoshioka, Hideaki Hakoshima, Yuichiro Matsuzaki, Yuuki Tokunaga, Yasunari Suzuki, and Suguru Endo. "Generalized Quantum Subspace Expansion". Physical Review Letters **129**, 020502 (2022).

[30] Bo Yang, Nobuyuki Yoshioka, Hiroyuki Harada, Shigeo Hakkaku, Yuuki Tokunaga, Hideaki Hakoshima, Kaoru Yamamoto, and Suguru Endo. "Resource-efficient generalized quantum subspace expansion". Phys. Rev. Appl. **23**, 054021 (2025).

[31] Yasuhiro Ohkura, Suguru Endo, Takahiko Satoh, Rodney Van Meter, and Nobuyuki Yoshioka. "Leveraging hardware-control imperfections for error mitigation via generalized quantum subspace" (2023). arXiv:2303.07660.

[32] Filip B. Maciejewski, Zoltán Zimborás, and Michał Oszmaniec. "Mitigation of readout noise in near-term quantum devices by classical post-processing based on detector tomography". Quantum **4**, 257 (2020).

[33] Andrew Arrasmith, Andrew Patterson, Alice Boughton, and Marco Paini. "Development and demonstration of an efficient readout error mitigation technique for use in nisq algorithms" (2023). arXiv:2303.17741.

[34] James Mills, Debasis Sadhukhan, and Elham Kashefi. "Simplifying errors by symmetry and randomisation" (2023). arXiv:2303.02712.

[35] Sergey Bravyi, Sarah Sheldon, Abhinav Kandala, David C. Mckay, and Jay M. Gambetta. "Mitigating measurement errors in multiqubit experiments". Physical Review A **103**, 042605 (2021).

[36] Suguru Endo, Zhenyu Cai, Simon C. Benjamin, and Xiao Yuan. "Hybrid Quantum-Classical Algorithms and Quantum Error Mitigation". Journal of the Physical Society of Japan **90**, 032001 (2021).

[37] Ying Li and Simon C. Benjamin. "Efficient Variational Quantum Simulator Incorporating Active Error Minimization". Physical Review X **7**, 021050 (2017).

[38] Daiqin Su, Robert Israel, Kunal Sharma, Haoyu Qi, Ish Dhand, and Kamil Brádler. "Error mitigation on a near-term quantum photonic device". Quantum **5**, 452 (2021).

[39] Walter Gautschi. "On inverses of Vandermonde and confluent Vandermonde matrices III". Numerische Mathematik **29**, 445–450 (1978).

[40] Giacomo De Palma, Milad Marvian, Cambyse Rouzé, and Daniel Stilck França. "Limitations of Variational Quantum Algorithms: A Quantum Optimal Transport Approach". PRX Quantum **4**, 010309 (2023).

[41] Ryuji Takagi, Hiroyasu Tajima, and Mile Gu. "Universal Sampling Lower Bounds for Quantum Error Mitigation". Physical Review Letters **131**, 210602 (2023).

[42] Yihui Quek, Daniel Stilck França, Sumeet Khatri, Johannes Jakob Meyer, and Jens Eisert. "Exponentially tighter bounds on limitations of quantum error mitigation". Nature Physics **20**, 1648–1658 (2024).

[43] Kento Tsubouchi, Takahiro Sagawa, and Nobuyuki Yoshioka. "Universal Cost Bound of Quantum Error Mitigation Based on Quantum Estimation Theory". Physical Review Letters **131**, 210601 (2023).

[44] Zoltán Zimborás, Bálint Koczor, Zoë Holmes, Elsi-Mari Borrelli, András Gilyén, Hsin-Yuan Huang, Zhenyu Cai, Antonio Acín, Leandro Aolita, Leonardo Banchi, et al. "Myths around quantum computation before full fault tolerance: What no-go theorems rule out and what they don't" (2025). arXiv:2501.05694.

[45] Youngseok Kim, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout Van Den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, et al. "Evidence for the utility of quantum computing before fault tolerance". Nature **618**, 500–505 (2023).

[46] Joseph Tindall, Matthew Fishman, E Miles Stoudenmire, and Dries Sels. "Efficient tensor network simulation of ibm's eagle kicked ising experiment". PRX Quantum **5**, 010308 (2024).

[47] Alexia Salavrakos, Tigran Sedrakyan, James Mills, Shane Mansfield, and Rawad Mezher. "Error-mitigated photonic quantum circuit born machine". Physical Review A **111**, L030401 (2025).

[48] Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. "Parameterized quantum circuits as machine learning models". Quantum Science and Technology **4**, 043001 (2019).

[49] Nicolas Maring, Andreas Fyrillas, Mathias Pont, Edouard Ivanov, Petr Stepanov, Nico Margaria, William Hease, Anton Pishchagin, Aristide Lemaître, Isabelle Sagnes, et al. "A versatile single-photon-based quantum computing platform". Nature Photonics **18**, 603–609 (2024).

[50] Donghwa Lee, Jinil Lee, Seongjin Hong, Hyang-Tag Lim, Young-Wook Cho, Sang-Wook Han, Hyundong Shin, Junaid ur Rehman, and Yong-Su Kim. "Error-mitigated photonic variational quantum eigensolver using a single-photon ququart". Optica **9**, 88–95 (2022).

[51] Nicolas Heurtel, Andreas Fyrillas, Grégoire de Gliniasty, Raphaël Le Bihan, Sébastien Malherbe, Marceau Pailhas, Eric Bertasi, Boris Bourdoncle, Pierre-Emmanuel Emeriau, Rawad Mezher, Luka Music, Nadia Belabas, Benoît Valiron, Pascale Senellart, Shane Mansfield, and Jean Senellart. "Perceval: A Software Platform for Discrete Variable Photonic Quantum Computing". Quantum **7**, 931 (2023).

[52] Beng Yee Gan, Daniel Leykam, and Dimitris G. Angelakis. "Fock State-enhanced Expressivity of Quantum Machine Learning Models". EPJ Quantum Technology **9**, 16 (2022).

[53] Rawad Mezher, Ana Filipa Carvalho, and Shane Mansfield. "Solving graph problems with single photons and linear optics". Physical Review A **108**, 032405 (2023).

[54] Adam Taylor, Gabriele Bressanini, Hyukjoon Kwon, and MS Kim. "Quantum error cancellation in photonic systems: Undoing photon losses". Physical Review A **110**, 022622 (2024).

[55] Sepehr Nezami. "Permanent of random matrices from representation theory: moments, numerics, concentration, and comments on hardness of boson-sampling" (2021). arXiv:2104.06423.

[56] Wassily Hoeffding. "The collected works of wassily hoeffding". Springer. (1994).

[57] Deepesh Singh, Gopikrishnan Muraleedharan, Boxiang Fu, Chen-Mou Cheng, Nicolas Roussy Newton, Peter Rohde, and Gavin Keith Brennen. "Proof-of-work consensus by quantum sampling". Quantum Science and Technology (2023).

[58] Michał Oszmaniec and Daniel J. Brod. "Classical simulation of photonic linear optics with lost particles". New Journal of Physics **20**, 092002 (2018).

[59] Thomas M. Stace and Sean D. Barrett. "Error correction and degeneracy in surface codes suffering loss". Physical Review A **81**, 022317 (2010).

[60] Raúl García-Patrón, Jelmer J. Renema, and Valery Shchesnovich. "Simulating boson sampling in lossy architectures". Quantum **3**, 169 (2019).

[61] Scott Aaronson and Daniel J. Brod. "BosonSampling with lost photons". Physical Review A **93**, 012335 (2016).

[62] Ross Berkowitz and Pat Devlin. "A stability result using the matrix norm to bound the permanent". Israel Journal of Mathematics **224**, 437–454 (2018).

[63] Zhengyan Lin and Zhidong Bai. "Probability Inequalities". Springer. Berlin, Heidelberg (2011).

[64] Rajendra Bhatia and Chandler Davis. "A Better Bound on the Variance". The American Mathematical Monthly **107**, 353–357 (2000).

[65] Kenneth Levenberg. "A method for the solution of certain non-linear problems in least squares". Quarterly of Applied Mathematics **2**, 164–168 (1944).

[66] Donald W. Marquardt. "An Algorithm for Least-Squares Estimation of Nonlinear Parameters". Journal of the Society for Industrial and Applied Mathematics **11**, 431–441 (1963).

[67] Scott Aaronson and Alex Arkhipov. "Bosonsampling is far from uniform". Quantum Information & Computation **14**, 1383–1423 (2014).

[68] Ulysse Chabaud, Damian Markham, and Adel Sohbi. "Quantum machine learning with adaptive linear optics". Quantum **5**, 496 (2021).

[69] Amara Katabarwa, Katerina Gratsea, Athena Caesura, and Peter D Johnson. "Early fault-tolerant quantum computing". PRX Quantum **5**, 020101 (2024).

[70] Herbert John Ryser. "Combinatorial Mathematics". American Mathematical Soc. (1963).

[71] J. L. W. V. Jensen. "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". Acta Mathematica **30**, 175–193 (1906).

[72] Scott Aaronson and Travis Hance. "Generalizing and derandomizing gurvits's approximation algorithm for the permanent" (2012). arXiv:1212.0025.

[73] Svante Janson. "A central limit theorem for m-dependent variables" (2021). arXiv:2108.12263.

[74] Benjamin Villalonga, Murphy Yuezhen Niu, Li Li, Hartmut Neven, John C. Platt, Vadim N. Smelyanskiy, and Sergio Boixo. "Efficient approximation of experimental gaussian boson sampling" (2022). arXiv:2109.11525.

Appendix

Table of Contents

9 Results on useful regimes of photon loss **33**
9.1 Proof of Lemma 6 . . . 33

10 Statistical error bounds for probability estimators **33**
10.1 Proof of Lemma 17 . . . 33

11 Interference deviation bounds **34**
11.1 Proof of Lemma 7 . . . 34
11.2 Proof of Thm. 8 . . . 35
11.3 Proof of Lemma 9 . . . 35
11.4 Proof of Thm. 10 . . . 36
11.5 Proofs of Lemmas 11 and 12 . . . 36
11.6 Proof of Thm. 13 . . . 38

12 Linear solving **39**
12.1 Bounding errors in linear solving . . . 39
12.2 Proof of Thm. 1 . . . 40
12.3 Linear solving with dependency . . . 42
12.4 Bias error bound with average dependency term d_k . . . 43
12.5 Performance guarantee inequality for linear solving with dependency . . . 44

13 Extrapolation **45**
13.1 Statistical error for average absolute deviation estimator $\tilde{D}_k$ . . . 45
13.2 Performance guarantee inequality for extrapolation with linear least squares . 47

14 Deterministic upper bound **52**
14.1 Proof of Thm. 14 . . . 52

15 Richardson extrapolation methods for photon loss mitigation **53**

16 Lyapunov bound **58**
16.1 Prospects of improving bounds on the interference terms . . . 58
16.2 Proof of Thm. 31 . . . 59

9 Results on useful regimes of photon loss

9.1 Proof of Lemma 6

We now prove Lemma 6 from the main text:

Let $k = n - r$, there is a classical algorithm running in time $O(2^{r-1}r(\binom{n}{n-r})^2)$ which exactly computes $\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} p(s_i^{n-k})$.

Proof. Each permanent $\mathsf{Per}(X_i)$ in $p(s_i^{n-k})$ can be computed exactly via Ryser's algorithm [70] in time $O(2^{r-1}r)$. Thus, each $p(s_i^{n-k})$ can be computed in $O(\binom{n}{n-r}2^{r-1}r)$ time. Further, we need $O(\binom{n}{n-r}^2 2^{r-1}r)$ time to compute all the $p(s_i^{n-k})$'s, and therefore to compute $\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} p(s_i^{n-k})$, by an iterative approach where we first set a variable $\mathsf{sum} = 0$, then iteratively add each computed $p(s_i^{n-k})$ to sum as soon as we compute it. □

10 Statistical error bounds for probability estimators

We now give upper bounds for the statistical error of the estimators of the ideal probabilities from postselection, and for the estimators of the recycled probabilities. These are used in later proofs.

10.1 Proof of Lemma 17

Lemma 17. *With probability at least $1 - 2e^{-\alpha^2}$ for $\alpha > 0$, the statistical error of each recycled probability, $\epsilon_{hoeff,\tilde{p}_R^k(s_l^n)}$, is upper bounded*

$$|\epsilon_{hoeff,\tilde{p}_R^k(s_l^n)}| \leq \alpha \frac{1}{\binom{m-n+k}{k}} \sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}.$$

Proof. The probability of losing k out of n photons is

$$\Pr(k) = \binom{n}{k}\eta^k(1-\eta)^{n-k},$$

and so the number of samples from $n-k$–photon outputs is $N_{tot,k} \approx \binom{n}{k}N_{tot}\eta^k(1-\eta)^{n-k}$, for $k \in \{0, \ldots, n\}$. There are $\binom{m}{n}$ recycled probabilities, and to guarantee the independence required for Hoeffding's inequality we use $N_{est,k} \approx \frac{\binom{n}{k}\eta^k(1-\eta)^{n-k}N_{tot}}{\binom{m}{n}}$ samples to estimate each $p_R^k(s_l^n)$. For each sample w ranging from 1 to $N_{est,k}$ assign the value 1 to a random variable $X_w \in \{0, 1\}$ if sample $w \in \mathcal{L}(s_l^n)$, and assign the value 0 to X_w otherwise. The estimator $\tilde{p}_R^k(s_i^n)$ of $p_R^k(s_i^n)$ is then

$$\tilde{p}_R^k(s_l^n) = \frac{\sum_w X_w}{\binom{m-n+k}{k}N_{est,k}},$$

where the $\binom{m-n+k}{k}^{-1}$ term is the normalisation factor detailed in eqn. 15. Finally, Hoeffding's inequality [56] gives with confidence at least $1 - 2e^{-\alpha^2}$ that

$$|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}| \leq \alpha \frac{1}{\binom{m-n+k}{k}} \sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}.$$

□

Note that, the above proof holds when using the collected samples to estimate *all* the probabilities, which is why we divided the k photon samples by $\binom{m}{n}$. If we wish to estimate

only one output probability, then we need not divide by $\binom{m}{n}$ and we can use all the $N_{tot,k}$ samples. Hoeffding's inequality gives us

$$|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}| \leq \alpha \frac{1}{\binom{m-n+k}{k}} \sqrt{\frac{1}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}, \tag{40}$$

with confidence at least $1 - 2e^{-\alpha^2}$.

11 Interference deviation bounds

11.1 Proof of Lemma 7

We now prove Lemma 7 from the main text:

Proof. Let $N'_k := \left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}$. For $k > 0$, $I_{s_l^n,k}$ is a sum of $N'_k := N_k - \binom{n}{k}$ terms of the form $p(s_j^n)\frac{1}{N_k}$. We can thus rewrite

$$I_{s_l^n,k} = \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{N'_k}.$$

From [67], when $m \gg n^2$, and U is Haar random, which corresponds to boson sampling unitaries, each $p(s_j^n) \approx \frac{|\mathsf{Per}(X_i)|^2}{m^n}$, where X_i is an $n \times n$ matrix of i.i.d. Gaussian random variables whose entries chosen from the complex normal distribution $\mathcal{N}_{\mathbb{C}}(0,1)$ of mean 0 and variance 1. $\mathsf{Per}(.)$ denotes the matrix permanent. Furthermore, from [67, 55], we know that

$$\mathrm{E}_{X \in \mathcal{G}_{n\times n}}(|\mathsf{Per}(X)|^2) = n!,$$

where $\mathrm{E}_{X \in \mathcal{G}_{n\times n}}(.)$ is the expectation value over the set of $n \times n$ Gaussian matrices with entries from $\mathcal{N}_{\mathbb{C}}(0,1)$. We can reexpress

$$I_{s_l^n,k} = \frac{1}{N'_k} \sum_i \frac{|\mathsf{Per}(X_i)|^2}{m^n}.$$

Now, from the linearity of the expected value

$$\mathrm{E}_{X \in \mathcal{G}_{n\times n}}(I_{s_l^n,k}) = \mathrm{E}_{X \in \mathcal{G}_{n\times n}}\left(\frac{1}{N'_k} \sum_i \frac{|\mathsf{Per}(X_i)|^2}{m^n}\right) = \frac{1}{m^n N'_k} \sum_i \mathrm{E}_{X_i \in \mathcal{G}_{n\times n}}(|\mathsf{Per}(X_i)|^2) = \frac{n!}{m^n},$$

where the rightmost part follows from the fact that $\mathrm{E}_{X_i \in \mathcal{G}_{n\times n}}(|\mathsf{Per}(X_i)|^2) = n!$, for all i.

Now $p_{unif} = \binom{m}{n}^{-1}$, and

$$\begin{aligned}\binom{m}{n}^{-1} &= \frac{n!(m-n)!}{m!} \\ &= \frac{n!}{(m)(m-1)\dots(m-n+1)} \\ &= \frac{n!}{(m^n + f(m,n))}\end{aligned}$$

with $f(m,n) \in O(m^{n-1})$. This leads to

$$\frac{n!}{m^n} = \binom{m}{n}^{-1} \frac{(m^n + f(m,n)))}{m^n},$$

which then becomes

$$\frac{n!}{m^n} = p_{unif}\big(1 + g(m)\big).$$

with $g(m) \in O(m^{-1})$, which was the result to be proved. □

11.2 Proof of Thm. 8

We now prove Lemma 8 from the main text:

The deviation of interference terms around p_{unif} for Haar random matrices is bounded

$$Pr\Big(\big|I_{s_l^n,k} - p_{unif}\big| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{np_{unif}^2}{\epsilon_{bias,s_l^n}^2},$$

where ϵ_{bias,s_l^n} is a positive real number.

Proof. Let $X := \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n)$, $M := \binom{n}{k}\left(\binom{m-n+k}{k} - 1\right)$. For simplicity, relabel $p(s_j^n) := p_i$, so that X can be rewritten as $X = \sum_{i=1,\ldots,M} p_i$ is a sum of, possibly dependent, random variables p_i and $\mathsf{E}_U(p_i) := \mathsf{E}_U(p) = \frac{n!}{m^n} \approx p_{unif}$, where $\mathsf{E}_U(p_i)$ is the expectation value over the Haar measure of the m-mode unitary group $\mathsf{U}(m)$. From [6], we know that this expectation value is given by $\frac{n!}{m^n}$ in the no-collision regime. Let $\sigma^2 = \mathsf{Var}(X) = \mathsf{E}_U(X^2) - (\mathsf{E}_U(X))^2$. Expanding out, we get that $\sigma^2 = M\mathsf{E}_U(p^2) - M(\mathsf{E}_U(p))^2 + 2\sum_i \sum_{j>i} \mathsf{Cov}(p_i, p_j)$, where $\mathsf{Cov}(p_i, p_j) = \mathsf{E}_U(p_i p_j) - \mathsf{E}_U(p_i)\mathsf{E}_U(p_j)$, and $\mathsf{E}_U(p^2) = \mathsf{E}_U(p_i^2) = \frac{(n+1)!n!}{m^{2n}} \approx (n+1)p_{unif}^2$, for all i, and this is from [6]. Using the Cauchy-Schwarz inequality,

$$|\mathsf{Cov}(p_i, p_j)| \leq \sqrt{\mathsf{Var}(p_i)\mathsf{Var}(p_j)} \leq \sqrt{(\mathsf{Var}(p))^2} \leq \mathsf{Var}(p) \leq \mathsf{E}_U(p^2) - (\mathsf{E}_U(p))^2$$

Thus,

$$\sigma^2 \leq M(\mathsf{E}_U(p^2) - (\mathsf{E}_U(p))^2 + 2\sum_i \sum_{j>i} |\mathsf{Cov}(p_i, p_j)| \leq M(\mathsf{E}_U(p^2) - (\mathsf{E}_U(p))^2) + M(M-1)(\mathsf{E}_U(p^2) - (\mathsf{E}_U(p))^2) \leq$$
$$M^2(\mathsf{E}_U(p^2) - (\mathsf{E}_U(p))^2) \leq M^2 np_{unif}^2.$$

Finally, using Chebyshev's inequality we have

$$Pr\Big(\big|X - \mathsf{E}_U(X)\big| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{\sigma^2}{\epsilon_{bias,s_l^n}^2},$$

and noting that $\mathsf{E}_U(X) = Mp_{unif} = \sum_i \mathsf{E}_U(p)$ from linearity of expectation value, then dividing both sides of $|X - \mathsf{E}_U(X)| \geq \epsilon_{bias,s_l^n}$ by M and redefining $\epsilon_{bias,s_l^n} := \frac{\epsilon_{bias,s_l^n}}{M}$, and finally using $\sigma^2 \leq M^2 np_{unif}^2$, we get

$$Pr\Big(\big|I_{s_l^n,k} - p_{unif}\big| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{np_{unif}^2}{\epsilon_{bias,s_l^n}^2}.$$

This completes the proof. □

11.3 Proof of Lemma 9

We now prove Lemma 9 from the main text:

$$\mathbf{E}_{D_R^k}\big(I_{s_l^n,k}\big) = p_{unif},$$

where $\mathbf{E}_{D_R^k}(.)$ denotes the expectation value over D_R^k.

Proof. Recalling that

$$p_R(s_l^n) = \frac{p(s_l^n)}{\binom{m-n+k}{k}} + \frac{N_k'}{N_k} I_{s_l^n,k},$$

and

$$I_{s_l^n,k} = \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}}.$$

$$
\begin{aligned}
\mathbf{E}_{D_R^k}\left(I_{s_l^n,k}\right) &= \frac{1}{\binom{m}{n}} \sum_{l=1}^{\binom{m}{n}} \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}} \\
&= \frac{1}{\binom{m}{n}\left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}} \sum_{l=1}^{\binom{m}{n}} \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \left(\sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) - p(s_l^n) \right) \\
&= \frac{1}{\binom{m}{n}\left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}} \left(\sum_{l=1}^{\binom{m}{n}} \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) - \sum_{l=1}^{\binom{m}{n}} \binom{n}{k} p(s_l^n) \right) \\
&= \frac{1}{\binom{m}{n}\left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}} \left(\sum_{l=1}^{\binom{m}{n}} \binom{m-n+k}{k} \binom{n}{k} p(s_l^n) - \sum_{l=1}^{\binom{m}{n}} \binom{n}{k} p(s_l^n) \right) \\
&= \frac{1}{\binom{m}{n}\left(\binom{m-n+k}{k} - 1\right)\binom{n}{k}} \left(\binom{m-n+k}{k} \binom{n}{k} - \binom{n}{k} \right) \\
&= p_{unif}.
\end{aligned}
$$

Where between the third and fourth lines we have used the cardinality of the sets $|\mathcal{L}(s_l^n)| = \binom{n}{k}$ and $|\mathcal{G}(s_i^{n-k})| = \binom{m-n+k}{k}$, and between the fourth and fifth lines that $\sum_{l=1}^{\binom{m}{n}} p(s_l^n) = 1$. □

11.4 Proof of Thm. 10

We now prove Thm. 10 from the main text:

The deviation of interference terms around p_{unif} for an arbitrary matrix is bounded

$$
\Pr\left(\left|I_{s_l^n,k} - p_{unif}\right| \geq \epsilon_{bias,s_l^n}\right) \leq \frac{p_{unif}}{\epsilon^2_{bias,s_l^n}},
$$

where ϵ_{bias,s_l^n} is a positive real number.

Proof. Let $X := I_{s_l^n,k}$ be a random variable defined over the uniform distribution of bit strings s_l^n that returns interference terms as values, $\mu := \mathsf{E}_{s_l^n}(X) = p_{unif} = \frac{1}{\binom{m}{n}}$, with $\mathsf{E}_{s_l^n}(.)$ denoting the expectation value of X over the uniform distribution of s_l^n, it is equal to p_{unif} from lemma 9. Let $\sigma^2 := \mathsf{Var}(X)$. From Chebyshev's inequality,

$$
Pr(|X - \mu| \geq \epsilon_{bias,s_l^n}) \leq \frac{\sigma^2}{\epsilon^2_{bias,s_l^n}}
$$

Let $M := \mathsf{max}_{s_l^n}(X)$, $m := \mathsf{min}_{s_l^n}(X)$, and note that $M \leq 1$ and $m \geq 0$. We now use the Bhatia-Davis inequality [64]

$$
\sigma^2 \leq (M - \mu)(\mu - m)
$$

with the upper and lower bounds on M and m to obtain

$$
\sigma^2 \leq (1 - p_{unif})(p_{unif}) \leq p_{unif}.
$$

Replacing this in Chebyshev's inequality completes the proof. □

11.5 Proofs of Lemmas 11 and 12

Both proofs will make use of Jensen's inequality [71]:

Theorem 18. (Jensen's inequality) *Let $f(x)$ be a convex function on (a, b) and suppose $a < x_1 \leq x_2 \leq \ldots \leq x_n < b$. Then*

$$\frac{f(x_1) + f(x_2) + \ldots + f(x_n)}{n} \geq f\Big(\frac{x_1 + x_2 + \ldots + x_n}{n}\Big).$$

Equality holds if, and only if, $x_1 = x_2 = \ldots = x_n$.

We now prove Lemma 11 from the main text:

The variance of the set of recycled probabilities is less than or equal to the variance of the set of ideal probabilities, that is

$$\mathsf{Var}\big(\{p(s_l^n)\}_l\big) \geq \mathsf{Var}\big(\{p_R^k(s_l^n)\}_l\big).$$

Proof. Let $\mu = p_{unif}$ and $f(x) = (x - \mu)^2$. Defining the variance of the ideal n output photon distribution as

$$\mathsf{Var}\big(\{p(s_l^n)\}_l\big) := \binom{m}{n}^{-1} \sum_l (p(s_l^n) - \mu)^2,$$

and the variance of the $n - k$ output photon recycled distribution as

$$\mathsf{Var}\big(\{p_R^k(s_l^n)\}_l\big) := \binom{m}{n}^{-1} \sum_l (p_R^k(s_l^n) - \mu)^2.$$

These definitions are equivalent to the variance of random variables uniformly distributed over the set of bit strings $\{s_l^n\}$ that return ideal n output photon probabilities in the first case, and $n - k$ output photon recycled probabilities in the second case. Recall that the recycled probabilities are defined by the expression

$$p_R^k(s_l^n) := \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) \frac{1}{\binom{m-n+k}{k}\binom{n}{k}}.$$

As $f(x)$ is concave, by applying Jensen's inequality it is true that

$$\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} \frac{f(p(s_j^n))}{\binom{m-n+k}{k}\binom{n}{k}} \geq f\Bigg(\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) \frac{1}{\binom{m-n+k}{k}\binom{n}{k}} \Bigg).$$

From which it follows

$$\binom{m}{n}^{-1} \sum_l \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} \frac{f(p(s_j^n))}{\binom{m-n+k}{k}\binom{n}{k}} \geq \binom{m}{n}^{-1} \sum_l f\Bigg(\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) \frac{1}{\binom{m-n+k}{k}\binom{n}{k}} \Bigg),$$

and after simplifying

$$\binom{m}{n}^{-1} \sum_l f(p(s_j^n)) \geq \binom{m}{n}^{-1} \sum_l f\Bigg(\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) \frac{1}{\binom{m-n+k}{k}\binom{n}{k}} \Bigg).$$

And this is just: $\mathsf{Var}\big(\{p(s_l^n)\}_l\big) \geq \mathsf{Var}\big(\{p_R^k(s_l^n)\}_l\big)$, the result to be shown. □

We now prove Lemma 12 from the main text:

The variance of the set of interference terms is less than or equal to the variance of the set of ideal probabilities, that is

$$\mathsf{Var}\big(\{p(s_l^n)\}_l\big) \geq \mathsf{Var}\big(\{I_{s_l^n,k}\}_l\big).$$

Proof. Let $\mu = p_{unif}$ and $f(x) = (x-\mu)^2$. Defining the variance of the ideal n output photon distribution as

$$\mathsf{Var}\big(\{p(s_l^n)\}_l\big) := \binom{m}{n}^{-1} \sum_l (p(s_l^n) - \mu)^2,$$

and the variance of the interference terms for the $n-k$ output photon recycled distribution as

$$\mathsf{Var}\big(\{I_{s_l^n,k}\}_l\big) := \binom{m}{n}^{-1} \sum_l \big(I_{s_l^n,k} - \mu\big)^2.$$

These definitions are equivalent to the variance of random variables uniformly distributed over the set of bit strings $\{s_l^n\}$ that return ideal n output photon probabilities in the first case, and interference terms from the $n-k$ output photon recycled distribution in the second case. Recall that the interference terms of a recycled distribution are defined by the expression

$$I_{s_l^n,k} = \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\big(\binom{m-n+k}{k} - 1\big)\binom{n}{k}}.$$

As $f(x)$ is concave, by applying Jensen's inequality it is true that

$$\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} \frac{f(p(s_j^n))}{\big(\binom{m-n+k}{k} - 1\big)\binom{n}{k}} \geq f\Bigg(\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\big(\binom{m-n+k}{k} - 1\big)\binom{n}{k}} \Bigg).$$

From which it follows

$$\begin{aligned} \binom{m}{n}^{-1} \sum_l \sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} & \frac{f(p(s_j^n))}{\big(\binom{m-n+k}{k} - 1\big)\binom{n}{k}} \\ & \geq \binom{m}{n}^{-1} \sum_l f\Bigg(\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k}), j \neq l} p(s_j^n) \frac{1}{\big(\binom{m-n+k}{k} - 1\big)\binom{n}{k}} \Bigg), \end{aligned}$$

and after simplifying

$$\binom{m}{n}^{-1} \sum_l f(p(s_j^n)) \geq \binom{m}{n}^{-1} \sum_l f\Bigg(\sum_{s_i^{n-k} \in \mathcal{L}(s_l^n)} \sum_{s_j^n \in \mathcal{G}(s_i^{n-k})} p(s_j^n) \frac{1}{\big(\binom{m-n+k}{k} - 1\big)\binom{n}{k}} \Bigg).$$

And this is just: $\mathsf{Var}\big(\{p(s_l^n)\}_l\big) \geq \mathsf{Var}\big(\{I_{s_l^n,k}\}_l)\}_l\big)$, the result to be shown. □

11.6 Proof of Thm. 13

We now prove Thm. 13 from the main text:

The deviation of interference terms around p_{unif} for an arbitrary matrix is bounded

$$\Pr\Big(\big|I_{s_l^n,k} - p_{unif}\big| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{p_{unif} p_{upper}}{\epsilon^2_{bias,s_l^n}} + \delta\Big(1 - \frac{p_{unif} p_{upper}}{\epsilon^2_{bias,s_l^n}}\Big),$$

where ϵ_{bias,s_l^n} *is a positive real number, and* p_{upper} *is an empirically computed upper bound on the largest probability of the ideal* n *output photon probability distribution with confidence* $1-\delta$.

Proof. Let $X := I_{s_l^n,k}$ be a random variable defined over the uniform distribution of bit strings s_l^n that returns interference terms as values. $\mathsf{E}_{s_l^n}(X)$ denotes the expectation value of random variable X over the uniform distribution of s_l^n, and let $\mu_X := \mathsf{E}_{s_l^n}(X) = p_{unif}$, the latter equality comes from lemma 9. Let $Y := p(s_l^n)$ be a random variable defined over the uniform distribution of bit strings s_l^n that returns n–photon output state probabilities as values, and $\mu_Y := \mathsf{E}_{s_l^n}(Y) = p_{unif}$,

where $\mathsf{E}_{s_l^n}(Y)$ denotes the expectation value of random variable Y over the uniform distribution of s_l^n. Using Chebyshev's inequality we have that

$$Pr(|X - \mu_X| \geq \epsilon_{bias,s_l^n}) \leq \frac{\mathsf{Var}(X)}{\epsilon^2_{bias,s_l^n}}$$

By lemma 12 we have that $\mathsf{Var}(X) \leq \mathsf{Var}(Y)$. Let $M := \mathsf{max}_{s_l^n}(Y)$, $m := \mathsf{min}_{s_l^n}(Y)$. An empirical upper bound on the largest probability on the largest probability, p_{upper}, may be computed from sample data such that $p_{upper} \geq M$. A Hoeffding inequality can be used to bound the statistical error of the empirical estimator of M, $\tilde{M}$, so that

$$|\epsilon_{hoeff,M}| \leq \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N_{est,M}}},$$

where $N_{est,M}$ is the number of samples used to compute the estimator $\tilde{M}$, and $1-\delta$ is the confidence. Defining $\epsilon^{max}_{hoeff,M} := \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2N_{est,M}}}$ and $p_{upper} := \tilde{M} + \epsilon^{max}_{hoeff,M}$. Then $p_{upper} > M$ with confidence $1 - \delta$. The smallest probability is lower bounded $m \geq 0$. We now use the Bhatia-Davis inequality [64]

$$\mathsf{Var}(Y) \leq (M - \mu_Y)(\mu_Y - m) = (M - p_{unif})(p_{unif} - m)$$

the upper and lower bounds on M and m then lead to

$$(M - p_{unif})(p_{unif} - m) \leq (p_{upper} - p_{unif})(p_{unif}) \leq p_{upper}p_{unif}.$$

The confidence that $|I_{s_l^n,k} - p_{unif}| < \epsilon_{bias,s_l^n}$ would be $1 - \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}}$, however this must be multiplied by the independent confidence of the statement $p_{upper} > M$, which is $1 - \delta$, to get an overall confidence of $1 - \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}} - \delta + \frac{p_{unif}p_{upper}\delta}{\epsilon^2_{bias,s_l^n}}$. This means the Chebyshev inequality becomes

$$\Pr\Big(\big|I_{s_l^n,k} - p_{unif}\big| \geq \epsilon_{bias,s_l^n}\Big) \leq \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}} + \delta\Big(1 - \frac{p_{unif}p_{upper}}{\epsilon^2_{bias,s_l^n}}\Big),$$

and the result follows. □

12 Linear solving

12.1 Bounding errors in linear solving

We first state and prove two results that we will use. These are upper bounds for the statistical error and the bias error of the mitigated value respectively, these we state as lemmas 19 and 20.

Lemma 19. $\epsilon_{mit,bias} \leq 3(\binom{m-n+k}{k} - 1)|\epsilon_{bias,s_l^n}|$.

Proof. Without statistical error the mitigated value may be written $p_{mit}(s_l^n) = \Big|\binom{m-n+k}{k}p_R(s_l^n) - \binom{m-n+k}{k}\frac{N'_k}{N_k}p_{unif}\Big|$. Let $\epsilon_{mit,bias} := |p_{mit}(s_l^n) - p(s_l^n)|$ and $A_k := \binom{m-n+k}{k}p_R(s_l^n) - \binom{m-n+k}{k}\frac{N'_k}{N_k}p_{unif}$, so that $\epsilon_{mit,bias} = ||A_k| - p(s_l^n)|$, then $p_{mit}(s_l^n) = |A_k|$, and $p(s_l^n) = A_k - \binom{m-n+k}{k}\frac{N'_k}{N_k}\epsilon_{bias,s_l^n}$ (see main text).

Consider the first possible case where $A_k < 0$, this means that

$$\begin{aligned}
\epsilon_{mit,bias} &= |-A_k - p(s_l^n)| \\
&= |-A_k - A_k + \binom{m-n+k}{k}\frac{N'_k}{N_k}\epsilon_{bias,s_l^n}| \\
&\leq 2|A_k| + \binom{m-n+k}{k}\frac{N'_k}{N_k}|\epsilon_{bias,s_l^n}|,
\end{aligned}$$

by triangle inequality. Since $p(s_l^n) \geq 0$ by definition, then $\binom{m-n+k}{k}\frac{N'_k}{N_k}|\epsilon_{bias,s_l^n}| = -\binom{m-n+k}{k}\epsilon_{bias,s_l^n} \geq -A_k$. Since $|A_k| = -A_k$, it follows that $|A_k| \leq \binom{m-n+k}{k}|\frac{N'_k}{N_k}\epsilon_{bias,s_l^n}|$, and therefore that $\epsilon_{mit,bias} \leq 2|A_k| + \binom{m-n+k}{k}\frac{N'_k}{N_k}|\epsilon_{bias,s_l^n}| \leq 3\binom{m-n+k}{k}\frac{N'_k}{N_k}|\epsilon_{bias,s_l^n}|$.

Now, consider the second possible case where $A_k \geq 0$, then $|A_k| = A_k$, and therefore $\epsilon_{mit,bias} = \binom{m-n+k}{k}\frac{N'_k}{N_k}|\epsilon_{bias,s_l^n}| \leq 3\binom{m-n+k}{k}\frac{N'_k}{N_k}|\epsilon_{bias,s_l^n}|$. This completes the proof. □

Lemma 20. $|\tilde{p}_{mit}(s_l^n) - p_{mit}(s_l^n)| \leq 3\binom{m-n+k}{k}|\epsilon_{hoeff,\tilde{p}_R^k(s_l^n)}|$.

Proof. First, note that the estimator of the mitigated value may be written $\tilde{p}_{mit} = |A_k + \epsilon_{mit,stat}|$, where A_k is defined previously, and $\epsilon_{mit,stat} = \binom{m-n+k}{k}\tilde{p}_R(s_l^n) - \binom{m-n+k}{k}p_R(s_l^n)$, and $|\epsilon_{mit,stat}| \leq \binom{m-n+k}{k}|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}|$. And therefore $|\tilde{p}_{mit} - p_{mit}| = ||A_k + \epsilon_{mit,stat}| - |A_k||$.

Consider the case where

$$|A_k + \epsilon_{mit,stat}| = A_k + \epsilon_{mit,stat}.$$

If $A_k > 0$, then $|A_k| = A_k$ and therefore $|\tilde{p}_{mit} - p_{mit}| = |\epsilon_{mit,stat}| \leq \binom{m-n+k}{k}|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}|$. If $A_k < 0$, then $A_k = -|A_k|$, and furthermore, since $A_k + \epsilon_{mit,stat} \geq 0$, then $\epsilon_{mit,stat} \geq |A_k|$. In this sub-case, we have that $|\tilde{p}_{mit} - p_{mit}| = |-2|A_k| - \epsilon_{mit,stat}| \leq 2|A_k| + |\epsilon_{mit,stat}| \leq 3\epsilon_{mit,stat} \leq 3\binom{m-n+k}{k}|\epsilon_{hoeff}|$.

Now, consider the case where

$$|A_k + \epsilon_{mit,stat}| = -A_k - \epsilon_{mit,stat}.$$

If $A_k < 0$, then $-A_k = |A_k|$, in this sub-case we have $|\tilde{p}_{mit} - p_{mit}| = |\epsilon_{mit,stat}| \leq \binom{m-n+k}{k}\epsilon_{hoeff}$. Now, if $A_k > 0$, then $-A_k = -|A_k|$ and furthermore, since $-A_k - \epsilon_{mit,stat} \geq 0$, then $-\epsilon_{mit,stat} \geq |A_k|$. In this sub-case, $|\tilde{p}_{mit} - p_{mit}| = |-2|A_k| - \epsilon_{mit,stat}| \leq 3|\epsilon_{mit,stat}| \leq 3\binom{m-n+k}{k}|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}|$.

In all possible cases,

$$|\tilde{p}_{mit} - p_{mit}| \leq 3\binom{m-n+k}{k}|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}|,$$

which was the result to be proved. □

12.2 Proof of Thm. 1

From Lemmas 19 and 20 we have that the overall error for linear solving recycling mitigation is given by

$$|\tilde{p}_{mit}(s_l^n) - p_{id}(s_l^n)| \leq 3\binom{m-n+k}{k}(|\epsilon_{bias,s_l^n}| + |\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}|). \tag{41}$$

Now rearranging Thm. 10 in terms of confidence parameter $\delta_{\text{bias}} \in (0,1)$, we get that

$$\Pr\left(|I_{s_l^n,k} - p_{unif}| \leq \sqrt{\frac{p_{\text{unif}}}{\delta_{\text{bias}}}}\right) \geq 1 - \delta_{\text{bias}}. \tag{42}$$

Now eqn. 40 states that

$$|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}| \leq \alpha\frac{1}{\binom{m-n+k}{k}}\sqrt{\frac{1}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}, \tag{43}$$

with confidence at least $1 - 2e^{-\alpha^2}$. Stated in the same form as eqn. 42, this is

$$\Pr\left(|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}| \leq \alpha\frac{1}{\binom{m-n+k}{k}}\sqrt{\frac{1}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}\right) \geq 1 - 2e^{-\alpha^2}. \tag{44}$$

Now applying union bound, with $|\epsilon_{bias,s_l^n}| := |I_{s_l^n,k} - p_{unif}|$, we have that

$$\Pr\left(|\epsilon_{bias,s_l^n}| + |\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}| \leq \sqrt{\frac{p_{\text{unif}}}{\delta_{\text{bias}}}} + \alpha \frac{1}{\binom{m-n+k}{k}} \sqrt{\frac{1}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}\right) \geq 1 - \delta_{\text{bias}} - 2e^{-\alpha^2}. \tag{45}$$

To obtain this expression in terms of a single confidence parameter we set $\delta_{bias} = 2e^{-\alpha^2}$, and the previous expression becomes

$$\Pr\left(|\epsilon_{bias,s_l^n}| + |\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}| \leq \frac{e^{\alpha^2/2}}{\sqrt{2}}\sqrt{p_{\text{unif}}} + \alpha \frac{1}{\binom{m-n+k}{k}} \sqrt{\frac{1}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}\right) \geq 1 - 4e^{-\alpha^2}. \tag{46}$$

We then get that

$$|\tilde{p}_{mit}(s_l^n) - p_{id}(s_l^n)| \leq 3\left[\frac{e^{\alpha^2/2}}{\sqrt{2}}\sqrt{p_{\text{unif}}}\binom{m-n+k}{k} + \alpha\sqrt{\frac{1}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}\right], \tag{47}$$

which holds with confidence $1 - 4e^{-\alpha^2}$.

To simplify this we will now make use of the well-known inequality that, for integers $a \geq b \geq 1$,

$$\left(\frac{a}{b}\right)^b \leq \binom{a}{b} \leq \left(\frac{ea}{b}\right)^b. \tag{48}$$

Using this inequality, we have that

$$\binom{m-n+k}{k} \leq \left(\frac{e(m-n+k)}{k}\right)^k, \tag{49}$$

and

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k}, \tag{50}$$

so that

$$\binom{n}{k}^{-1} \leq \left(\frac{k}{n}\right)^k. \tag{51}$$

This may be used to then derive the following:

$$|\tilde{p}_{mit}(s_l^n) - p_{id}(s_l^n)| \leq 3\left[\frac{e^{\alpha^2/2}}{\sqrt{2}}\sqrt{p_{\text{unif}}}\left(\frac{e(m-n+k)}{k}\right)^k + \alpha\left(\frac{k}{n}\right)^{k/2}\sqrt{\frac{1}{(1-\eta)^{n-k}\eta^k N_{tot}}}\right], \tag{52}$$

which holds with confidence $1 - 4e^{-\alpha^2}$. We now set $\delta = 4e^{-\alpha^2}$. Then $\alpha = \sqrt{\ln(4/\delta)}$ and $\frac{e^{\alpha^2/2}}{\sqrt{2}} = \sqrt{\frac{2}{\delta}}$. Substituting this, we then get the following bound

$$\left|\tilde{p}_{\text{mit}}(s_l^n) - p_{\text{id}}(s_l^n)\right| \leq 3\left[\sqrt{\frac{2p_{\text{unif}}}{\delta}}\left(\frac{e(m-n+k)}{k}\right)^k + \sqrt{\ln(4/\delta)}\left(\frac{k}{n}\right)^{k/2}\sqrt{\frac{1}{(1-\eta)^{n-k}\eta^k N_{\text{tot}}}}\right], \tag{53}$$

which holds with confidence $1 - \delta$. Finally, since $p_{\text{unif}} = \binom{m}{n}^{-1}$ and

$$\binom{m}{n} \geq \left(\frac{m}{n}\right)^n, \tag{54}$$

we have

$$p_{\text{unif}} \leq \left(\frac{n}{m}\right)^n, \qquad \sqrt{p_{\text{unif}}} \leq \left(\frac{n}{m}\right)^{n/2}. \tag{55}$$

Substituting this bound yields the expression for $f(n, m, k, \eta, N_{tot})$ in Theorem 1.

12.3 Linear solving with dependency

We now consider the case where there is positive correlation of the interference to ideal probabilities within the recycled probabilities. This effect can be modelled as a linear dependence of interference terms on respective ideal probabilities. For a given recycled probability, a dependency term $d_k(s_l^n)$ can be used to express the interference term as a linear function of $p(s_l^n)$ and p_{unif}. Then

$$\big(1 - d_k(s_l^n)\big)p_{unif} + d_k(s_l^n)p(s_l^n) = I_{s_l^n,k}$$

defines the value $d_k(s_l^n)$, where $k \leq n-1$. For each interference term, $I_{s_l^n,k}$, there is a corresponding dependency term, $d_k(s_l^n)$. The dependency term, $d_k(s_l^n)$, encodes the dependence of interference term $I_{s_l^n,k}$ on ideal probability $p(s_l^n)$ for the recycled probability $p_R^k(s_l^n)$. The set of dependency terms is then denoted $\{d_k(s_l^n)\}_l$.

The original linear solving method involves approximating the interference term as p_{unif}, and solving to find the mitigated probability. Now we propose first extracting an average dependency term from the distribution, which we will call the *average dependency term* d_k, and then approximating the interference terms as $\big(1 - d_k\big)p_{unif} + d_k p(s_l^n)$ and solving as before. The motivation for this variation on the original protocol is to improve mitigation performance by capturing the enhanced ideal probability signal caused by this correlation behaviour. In the original solving method, excluding statistical error, the recycled probability is decomposed as

$$p_R(s_l^n) = \frac{p(s_l^n)}{\binom{m-n+k}{k}} + \frac{N_k'}{N_k}(p_{unif} + \epsilon_{bias,s_l^n}),$$

where ϵ_{bias,s_l^n} is the bias error from approximating $I_{s_l^n,k}$ as $\frac{N_k'}{N_k}p_{unif}$. If the interference term $I_{s_l^n,k}$ is instead approximated as $\frac{N_k'}{N_k}\Big(\big(1 - d_k\big)p_{unif} + d_k p(s_l^n)\Big)$, then the expression becomes

$$\begin{aligned} p_R^k(s_l^n) &= \frac{p(s_l^n)}{\binom{m-n+k}{k}} + \frac{N_k'}{N_k}\Big(\big((1 - d_k)p_{unif} + d_k p(s_l^n)\big) + \epsilon_b\Big) \\ &= p(s_l^n)\left(\frac{1}{\binom{m-n+k}{k}} + \frac{N_k'}{N_k}d_k\right) + \frac{N_k'}{N_k}\big((1 - d_k)p_{unif} + \epsilon_b\big), \end{aligned}$$

where ϵ_b is the bias error in the interference terms away from the linear model. From the positivity of the interference terms

$$\epsilon_b \geq -\big((1 - d_k)p_{unif} + d_k p(s_l^n)\big),$$

where $k \leq n-1$. One factor motivating the inclusion of a dependency term is that positive correlation of the interference terms with the ideal probabilities means that rather than the signal of the ideal probability having magnitude $\frac{1}{\binom{m-n+k}{k}}$, its magnitude is instead $\frac{1}{\binom{m-n+k}{k}} + \frac{N_k'}{N_k}d_k$. And this stronger signal may be used to improve mitigation performance. We now show how to compute the dependency term d_k from the absolute average deviation. Then compute bounds on the bias error and statistical error for the use of linear dependency in linear solving, which are then used to provide performance guarantees relative to postselection.

12.3.1 Computing the average dependency term d_k

One consequence of a general correlation of interference terms with ideal probabilities within recycled probabilities is that $D_k^{\text{no dep.}} \leq D_k$. The average dependency term d_k is a weighted average of dependency terms, it can be computed from $D_k^{\text{no dep.}}$ and D_k. We know by definition that

$$D_0 = \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n \in \boldsymbol{S}} |p(s_l^n) - p_{unif}|.$$

For the moment ignoring statistical error, the dependency factor can be calculated from the average distance of probabilities from the uniform probability for the $n-k$–photon recycled distribution

$$\begin{aligned}
D_k &= \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\frac{\binom{n}{k}p(s_l^n)}{N_k}+\frac{N_k'}{N_k}\big((1-d_k(s_l^n))p_{unif}+d_k(s_l^n)p(s_l^n)\big)-p_{unif}\right| \\
&= \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\frac{\binom{n}{k}p(s_l^n)}{N_k}+\frac{N_k'}{N_k}p_{unif}-\frac{N_k'}{N_k}d_k(s_l^n)p_{unif}+\frac{N_k'}{N_k}d_k(s_l^n)p(s_l^n)-p_{unif}\right| \\
&= \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\left(\frac{\binom{n}{k}}{N_k}\right)\Big(p(s_l^n)-p_{unif}\Big)+d_k(s_l^n)\frac{N_k'}{N_k}\Big(p(s_l^n)-p_{unif}\Big)\right| \\
&= \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\left(\frac{\binom{n}{k}}{N_k}+d_k(s_l^n)\frac{N_k'}{N_k}\right)\Big(p(s_l^n)-p_{unif}\Big)\right|
\end{aligned}$$

The average dependency term, d_k, for the recycled distribution is defined by the expression

$$\left(\frac{\binom{n}{k}}{N_k}+d_k\frac{N_k'}{N_k}\right)\frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\Big(p(s_l^n)-p_{unif}\Big)\right| = \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\left(\frac{\binom{n}{k}}{N_k}+d_k\big(p(s_l^n)\big)\frac{N_k'}{N_k}\right)\Big(p(s_l^n)-p_{unif}\Big)\right|.$$

This can then be used to rewrite the previous expression in terms of the dependency

$$\begin{aligned}
D_k &= \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\left(\frac{\binom{n}{k}}{N_k}+d_k(s_l^n)\frac{N_k'}{N_k}\right)\Big(p(s_l^n)-p_{unif}\Big)\right| \\
&= \left(\frac{\binom{n}{k}}{N_k}+d_k\frac{N_k'}{N_k}\right)\frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\Big(p(s_l^n)-p_{unif}\Big)\right| \\
&= \left(1+d_k\left(\binom{m-n+k}{k}-1\right)\right)\frac{1}{\binom{m-n+k}{k}}\frac{1}{|\boldsymbol{S}|}\sum_{s_l^n\in\boldsymbol{S}}\left|\Big(p(s_l^n)-p_{unif}\Big)\right| \\
&= \left(\frac{1+d_k\binom{m-n+k}{k}-d_k}{\binom{m-n+k}{k}}\right)D_0 \\
&= \left(\frac{1}{\binom{m-n+k}{k}}+d_k\frac{\binom{m-n+k}{k}-1}{\binom{m-n+k}{k}}\right)D_0.
\end{aligned}$$

If there is no correlation of interference terms with ideal probabilities then $d_k=0$, and then

$$D_k^{d_k=0} = \frac{1}{\binom{m-n+k}{k}}D_0.$$

Meaning that if there is no dependence the average absolute deviation decays with k proportionally to $\frac{1}{\binom{m-n+k}{k}}$. Whereas with dependency the decay is proportional to $\left(\frac{1+d_k\binom{m-n+k}{k}-d_k}{\binom{m-n+k}{k}}\right)$. The average dependency term may be directly computed from the previous expression as

$$d_k = \frac{1}{\binom{m-n+k}{k}-1}\left(\frac{\binom{m-n+k}{k}D_k}{D_0}-\frac{1}{\binom{m-n+k}{k}}\right).$$

12.4 Bias error bound with average dependency term d_k

We now prove Lemma 21, which will be used in the next section.

Lemma 21. *The bias error from substituting $I_{s_l^n,k}$ with $\big(1-d_k\big)p_{unif}+d_kp(s_l^n)$ in the recycled probabilities is upper bounded*

$$Pr\Big[\big|I_{s_l^n,k}-\big((1-d_k)p_{unif}+d_kp(s_l^n)\big)\big|\geq 2\epsilon_{bias,s_l^n}\Big]\leq\frac{2p_{unif}}{\epsilon^2_{bias,s_l^n}}.$$

Proof. Rather than bounding $I_{s_l^n,k}$ from p_{unif}, we would now like to bound deviation of $I_{s_l^n,k}$ from $\big(1-d_k\big)p_{unif}+d_k p(s_l^n)$. As it is required that $1 \geq d_k \geq 0$, terms of the form $\big(\big(1-d_k\big)p_{unif}+d_k p(s_l^n)\big)$ are bounded

$$|p(s_l^n) - \big((1-d_k)p_{unif} + d_k p(s_l^n)\big)| \geq |\big((1-d_k)p_{unif} + d_k p(s_l^n)\big) - p_{unif}|,$$

or

$$|p(s_l^n) - \big((1-d_k)p_{unif} + d_k p(s_l^n)\big)| \leq |\big((1-d_k)p_{unif} + d_k p(s_l^n)\big) - p_{unif}|,$$

with both statements true only with equidistance. The mean of the set $\{p(s_l^n)\}_l$ is p_{unif}, and so also the mean of the set $\{\big(1-d_k\big)p_{unif} + d_k p(s_l^n)\}_l$ is p_{unif}. If $d_k = 0$ then $\mathsf{Var}(\{\big(1-d_k\big)p_{unif} + d_k p(s_l^n)\}_l) = 0$, while if $d_k = 1$ then $\mathsf{Var}(\{\big(1-d_k\big)p_{unif} + d_k p(s_l^n)\}_l) = \mathsf{Var}(\{p(s_l^n)\}_l)$. As $d_k \leq 1$ the variance of the set $\{\big(1-d_k\big)p_{unif} + d_k p(s_l^n)\}_l$ is upper bounded by the variance of $\{p(s_l^n)\}_l$, so $\mathsf{Var}\big((1-d_k)p_{unif} + d_k p(s_l^n)\big) \leq p_{unif}$, where we have used the Bhatia-Davis inequality. We can then bound the distance of $\big((1-d_k)p_{unif} + d_k p(s_l^n)\big)$ from p_{unif} using Chebyshev's inequality so that

$$\Pr\left[\big|\big((1-d_k)p_{unif} + d_k p(s_l^n) - p_{unif}\big)\big| \geq \epsilon_{bias,s_l^n}\right] \leq \frac{p_{unif}}{\epsilon^2_{bias,s_l^n}}.$$

And in reverse form this inequality is

$$\Pr\left[\big|\big((1-d_k)p_{unif} + d_k p(s_l^n) - p_{unif}\big)\big| < \epsilon_{bias,s_l^n}\right] \geq 1 - \frac{p_{unif}}{\epsilon^2_{bias,s_l^n}}. \tag{56}$$

The Thm. 10 bound on the bias of the interference term for arbitrary matrices is

$$\Pr\left[\big|I_{s_l^n,k} - p_{unif}\big| \geq \epsilon_{bias,s_l^n}\right] \leq \frac{p_{unif}}{\epsilon^2_{bias,s_l^n}}.$$

Which in reverse form is

$$\Pr\left[\big|I_{s_l^n,k} - p_{unif}\big| < \epsilon_{bias,s_l^n}\right] \geq 1 - \frac{p_{unif}}{\epsilon^2_{bias,s_l^n}}. \tag{57}$$

Now, we use triangle inequality to combine the reverse forms of the above two inequalities, so that

$$\begin{aligned} |I_{s_l^n,k} - \big((1-d_k)p_{unif} + d_k p(s_l^n)\big)| &\leq |I_{s_l^n,k} - p_{unif}| + |-\big((1-d_k)p_{unif} + d_k p(s_l^n)\big) + p_{unif}| \\ &\leq 2\epsilon_{bias,s_l^n}. \end{aligned} \tag{58}$$

Assuming independence of the inequalities in eqn. 56 and eqn. 57, the the inequality in eqn. 58 holds with confidence $(1-\frac{p_{unif}}{\epsilon^2_{bias,s_l^n}})^2$. Using a Bernoulli approximation $(1-\frac{p_{unif}}{\epsilon^2_{bias,s_l^n}})^2 \approx 1-2\frac{p_{unif}}{\epsilon^2_{bias,s_l^n}}$, and the result follows.

□

12.5 Performance guarantee inequality for linear solving with dependency

We now derive the condition for when recycling mitigation with linear solving with dependency can outperform postselection.

Theorem 22. *The condition*

$$\sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}} + C\left(\binom{m-n+k}{k} - 1\right)\cdot \epsilon_{bias,s_l^n} \leq \sqrt{\frac{\binom{m}{n}}{(1-\eta)^n N_{tot}}},$$

with $C > 0$ defines a sampling regime where the sum of the worst-case statistical error and bias error of linear solving with dependency recycling mitigation is less than the worst-case statistical error of postselection.

Proof. The estimator of the recycled probability may be rewritten in terms of the dependency term d_k as

$$\begin{aligned} p_R^k(s_l^n) &= \frac{p(s_l^n)}{\binom{m-n+k}{k}} + \frac{N'_k}{N_k}\Big(\big((1-d_k-\epsilon_{\text{hoeff},d_k})p_{unif} + (d_k+\epsilon_{\text{hoeff},d_k})p(s_l^n)\big) + \epsilon_b\Big) + \epsilon_{\text{hoeff},p_R^k} \\ &= p(s_l^n)\left(\frac{1}{\binom{m-n+k}{k}} + \frac{N'_k}{N_k}(d_k+\epsilon_{\text{hoeff},d_k})\right) + \frac{N'_k}{N_k}\big((1-d_k-\epsilon_{\text{hoeff},d_k})p_{unif} + \epsilon_{bias,s_l^n}\big) + \epsilon_{\text{hoeff},p_R^k}. \end{aligned}$$

Solving to find the mitigated value in the ideal case one obtains

$$p_{mit,dep}(s_l^n) = \left|\frac{p_R^k(s_l^n) + \frac{N'_k}{N_k}\big((-1+d_k)p_{unif}\big)}{\frac{1}{\binom{m-n+k}{k}} + \frac{N'_k}{N_k}d_k}\right|.$$

However, the estimator of the mitigated value including error is

$$\tilde{p}_{mit,dep}(s_l^n) = \frac{p_R^k(s_l^n) - \frac{N'_k}{N_k}\big((1-(d_k+\epsilon_{\text{hoeff},d_k}))p_{unif} + \epsilon_{bias,s_l^n}\big) + \epsilon_{\text{hoeff},p_R^k}}{\frac{1}{\binom{m-n+k}{k}} + \frac{N'_k}{N_k}(d_k+\epsilon_{\text{hoeff},d_k})}.$$

As, by definition, $1 \geq d_k \geq 0$, this means that $-d_k \leq \epsilon_{\text{hoeff},d_k} \leq 1$. And the value of $\epsilon_{\text{hoeff},d_k}$ that maximally increases the error of the above quotient is then $\epsilon_{\text{hoeff},d_k} = -d_k$, this substitution removes the d_k terms. From this point we can use Lemma 19 and Lemma 20, with the one difference in the latter that the bias error ϵ_{bias,s_l^n} from Lemma 21 is used instead of ϵ_{bias,s_l^n}. And then the total error of the mitigated probability, $\epsilon_{p_{mit,dep}(s_l^n)}$, can be upper bounded

$$\left|\epsilon_{p_{mit,dep}(s_l^n)}\right| \leq 3\binom{m-n+k}{k}\left(\left|\epsilon_{\text{hoeff},p_R^k}\right| + \frac{N'_k}{N_k}\left|\epsilon_{bias,s_l^n}\right|\right).$$

Now we use that

$$|\epsilon_{\text{hoeff},p_R^k}| \leq D\frac{1}{\binom{m-n+k}{k}}\sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}},$$

for some $D > 1$ large enough so the inequality holds with high confidence. So that the overall error for linear solving with dependency is upper bounded

$$\left|\epsilon_{p_{mit,dep}(s_l^n)}\right| \leq \binom{m-n+k}{k}\left(D\frac{1}{\binom{m-n+k}{k}}\sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}} + \frac{2N'_k}{N_k}\cdot\epsilon_{bias,s_l^n}\right).$$

And so the condition to beat postselection is

$$\binom{m-n+k}{k}\left(\frac{1}{\binom{m-n+k}{k}}\sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}} + \frac{2}{D}\frac{N'_k}{N_k}\cdot\epsilon_{bias,s_l^n}\right) \leq \sqrt{\frac{\binom{m}{n}}{(1-\eta)^n N_{tot}}},$$

that is rearranged for the result, where we have the same constant D for Hoeffding's inequality for postselection. □

13 Extrapolation

13.1 Statistical error for average absolute deviation estimator $\tilde{D}_k$

The following lemma upper bounds the statistical error of the empirically computed absolute average deviation terms $\{\tilde{D}_k\}_{k=1}^K$, and is used in deriving the condition for linear extrapolation to outperform postselection in the next section.

Lemma 23. *The statistical error of the average absolute deviation estimator $\tilde{D}_k$ is upper bounded*

$$\left|\epsilon_{hoeff,\tilde{D}_k}\right| \in O\left(\left(\frac{1}{\binom{m-n+k}{k}^2\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}\right)^{1/4}\right). \tag{59}$$

Proof. Let X be a random variable defined over the uniform distribution of bit strings s_l^n that returns terms from the set $\{\tilde{p}_R^k(s_l^n)\}_l$ as values, $\mu := \mathsf{E}_{s_l^n}(X) = p_{unif} = \frac{1}{\binom{m}{n}}$, with $\mathsf{E}_{s_l^n}(.)$ denoting the expectation value of X over the uniform distribution of s_l^n, it is equal to p_{unif}. The average absolute deviation estimator, denoted $\{\tilde{D}_k\}_{k=0}^n$, is defined

$$\tilde{D}_k := \binom{m}{n}^{-1} \sum_{s_l^n \in \boldsymbol{S}} \left|\tilde{p}_R^k(s_l^n) - p_{unif}\right| = D_k + \epsilon_{\text{hoeff},\tilde{D}_k},$$

where $\epsilon_{\text{hoeff},\tilde{D}_k}$ is the absolute average deviation statistical error. Let $M := \mathsf{max}_{s_l^n}(X)$, $m := \mathsf{min}_{s_l^n}(X)$, and note that $M \leq 1$ and $m \geq 0$. We can apply the Bhatia-Davis inequality [64]

$$\mathsf{Var}(\{X\}) \leq (M - \mu)(\mu - m),$$

with an upper bound of $M = 1$, a lower bound of $m = 0$, and the expected value $\mu = p_{unif}$, to obtain an upper bound on the variance of the random variable X so that

$$\begin{aligned} \mathsf{Var}(X) &\leq (1 - p_{unif})(p_{unif}) \\ &\leq p_{unif}. \end{aligned} \tag{60}$$

Now, by Jensen's inequality [71] we know that $\mathsf{E}_{s_l^n}(|X - \mu|^2) \geq (\mathsf{E}_{s_l^n}|X - \mu|)^2$, therefore

$$\mathsf{Var}(X) \geq \left(\binom{m}{n}^{-1} \sum_{s_l^n \in \boldsymbol{S}} \left|\tilde{p}_R^k(s_l^n) - p_{unif}\right|\right)^2,$$

and as both sides of this inequality are positive and the square root operation is a monotonically increasing function for positive reals

$$\mathsf{Var}(X)^{1/2} \geq \binom{m}{n}^{-1} \sum_{s_l^n \in \boldsymbol{S}} \left|\tilde{p}_R^k(s_l^n) - p_{unif}\right|.$$

And using the result from eqn. 60 we have

$$\begin{aligned} p_{unif}^{1/2} &\geq \binom{m}{n}^{-1} \sum_{s_l^n \in \boldsymbol{S}} \left|\tilde{p}_R^k(s_l^n) - p_{unif}\right| \\ &\geq D_k + \epsilon_{\text{hoeff},\tilde{D}_k}. \end{aligned}$$

Now because $\tilde{D}_k \geq 0$ then $\epsilon_{\text{hoeff},\tilde{D}_k} \geq -D_k \geq -p_{unif}^{1/2}$. Also $p_{unif}^{1/2} \geq D_k$ and $p_{unif}^{1/2} \geq D_k + \epsilon_{\text{hoeff},\tilde{D}_k}$ mean that $2(p_{unif}^{1/2} - D_k) \geq \epsilon_{\text{hoeff},\tilde{D}_k}$. And as $D_k \geq 0$ this means $2p_{unif}^{1/2} \geq \epsilon_{\text{hoeff},\tilde{D}_k}$. Then we have that $2p_{unif}^{1/2} \geq \epsilon_{\text{hoeff},\tilde{D}_k} \geq -p_{unif}^{1/2}$ which gives

$$|\epsilon_{\text{hoeff},\tilde{D}_k}| \leq 2p_{unif}^{1/2}. \tag{61}$$

The absolute error for $\tilde{D}_k$ can also be upper bounded in terms of the statistical error of the recycled probabilities

$$\begin{aligned} |\epsilon_{\text{hoeff},\tilde{D}_k}| &= |\tilde{D}_k - D_k| \\ &= \left|\binom{m}{n}^{-1} \left(\sum_{s_l^n \in \boldsymbol{S}} \left|p_R^k(s_l^n) + \epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)} - p_{unif}\right| - \sum_{s_l^n \in \boldsymbol{S}} \left|p_R^k(s_l^n) - p_{unif}\right|\right)\right| \\ &\leq \binom{m}{n}^{-1} \left(\sum_{s_l^n \in \boldsymbol{S}} \left|\left|p_R^k(s_l^n) + \epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)} - p_{unif}\right| - \left|p_R^k(s_l^n) - p_{unif}\right|\right|\right) \\ &\leq \binom{m}{n}^{-1} \sum_{s_l^n \in \boldsymbol{S}} \left|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}\right|. \end{aligned} \tag{62}$$

Where the inequalities follow by using a triangle, then a reverse triangle inequality. All terms in the inequalities from eqns. 61 and 62 are positive reals, and so the inequalities may be combined to give

$$|\epsilon_{\text{hoeff},\tilde{D}_k}|^2 \leq 2p_{unif}^{1/2}\binom{m}{n}^{-1}\sum_{s_l^n \in \boldsymbol{S}}\left|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}\right|.$$

We now define $\mathcal{E}_{\text{hoeff},\tilde{p}_R^k}$ as the magnitude of the upper bound for the statistical error $|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}|$ which holds with high confidence from Hoeffdings inequality.

$$\mathcal{E}_{\text{hoeff},\tilde{p}_R^k} \in O\left(\frac{1}{\binom{m-n+k}{k}}\sqrt{\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}}\right).$$

As $|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}| \leq \mathcal{E}_{\text{hoeff},\tilde{p}_R^k}$ we can now write

$$|\epsilon_{\text{hoeff},\tilde{D}_k}|^2 \leq 2p_{unif}^{1/2}\mathcal{E}_{\text{hoeff},\tilde{p}_R^k},$$

and, again using that the square root operation is a monotonically increasing function for positive reals, it follows that

$$\begin{aligned}
|\epsilon_{\text{hoeff},\tilde{D}_k}| &\leq (2\mathcal{E}_{\text{hoeff},\tilde{p}_R^k})^{1/2}p_{unif}^{1/4} \\
&\in O\big(\mathcal{E}_{\text{hoeff},\tilde{p}_R^k}^{1/2}p_{unif}^{1/4}\big) \\
&\in O\left(\left(\frac{1}{\binom{m-n+k}{k}}\right)^{1/2}\left(\frac{\binom{m}{n}}{\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}\right)^{1/4}\left(\frac{1}{\binom{m}{n}}\right)^{1/4}\right) \\
&\in O\left(\left(\frac{1}{\binom{m-n+k}{k}^2\binom{n}{k}(1-\eta)^{n-k}\eta^k N_{tot}}\right)^{1/4}\right).
\end{aligned}$$

Which was the result to be proved. □

13.2 Performance guarantee inequality for extrapolation with linear least squares

Before giving a proof of Thm. 26 we first state and prove two results that we will use. These are an upper bound for the statistical error for the gradient parameter. And an upper bound on the bias error introduce by using the average gradient parameter rather than the ideal gradient parameter. These we state as Lemmas 24 and 25.

Lemma 24. *The average gradient parameter statistical error $\epsilon_{hoeff,g}$ is such that*

$$\left|\epsilon_{hoeff,g}\right| \in O\left(\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4}\right). \tag{63}$$

Proof. The bias error is $|g_{ideal-g_{avg}}|$. First note that using the linear least squares method the average gradient parameter estimator may be written

$$\tilde{g}_{avg} = \frac{3}{2n_d+1}\tilde{D}_0 - \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}\tilde{D}_i x_i,$$

and the average gradient parameter without statistical error

$$g_{avg} = \frac{3}{2n_d+1}D_0 - \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}D_i x_i.$$

The statistical error for the gradient term is then

$$
\begin{aligned}
\left|\epsilon_{\text{hoeff},g}\right| &= |\tilde{g}_{avg} - g_{avg}| \\
&= \Big|\frac{3}{2n_d+1}\tilde{D}_0 - \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}\tilde{D}_i x_i - \frac{3}{2n_d+1}D_0 + \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}D_i x_i\Big| \\
&= \Big|\frac{3}{2n_d+1}(D_0+\epsilon_{\text{hoeff},\tilde{D}_0}) - \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}(D_i+\epsilon_{\text{hoeff},\tilde{D}_i})x_i - \frac{3}{2n_d+1}D_0 \\
&\qquad + \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}D_i x_i\Big| \\
&= \Big|\frac{3}{2n_d+1}\epsilon_{\text{hoeff},\tilde{D}_0} + \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}\epsilon_{\text{hoeff},\tilde{D}_i}x_i\Big| \\
&\leq \frac{3}{2n_d+1}\left|\epsilon_{\text{hoeff},\tilde{D}_0}\right| + \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}\left|\epsilon_{\text{hoeff},\tilde{D}_i}\right|x_i \\
&\leq \frac{3}{2n_d+1}A\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4} + B\frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4}x_i \\
&\in O\left(\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4}\right),
\end{aligned}
$$

for $A, B > 0$. Where a triangle inequality was used for the fourth line, and that $\left|\epsilon_{\text{hoeff},\tilde{D}_i} = B\big(\frac{1}{(1-\eta)^n N_{tot}}\big)^{1/4}\right.$ for $i \in \{0,\ldots,n_d\}$ from lemma 23 was used to get the sixth line.

□

Lemma 25. *The bias error $\epsilon_{bias,g}$ from substituting the ideal gradient parameter, g_{ideal}, for the average gradient parameter, g_{avg}, is upper bounded*

$$\left|\epsilon_{bias,g}\right| \in O\left(\binom{m}{n}^{-1/3}\right), \tag{64}$$

with confidence $1 - O(m^{-n/3})$.

Proof. The absolute bias error from the use of the average gradient rather than the ideal gradient in the second iteration of linear least squares is $\epsilon_{\text{bias},g} = |g_{ideal} - g_{avg}|$. The ideal α_{ideal,s_l^n} term that will result in the ideal probability being produced by the least squares method may be written

$$\alpha_{ideal,s_l^n} = \frac{n_d+1}{2}g_{ideal} + \frac{1}{n_d}\sum_{i=1}^{n_d}(p_R^k(s_l^n) - p_{unif}), \tag{65}$$

and because we are considering the ideal case it is true that $p(s_l^n) = \alpha_{ideal,s_l^n} + p_{unif}$. From $\alpha_{ideal,s_l^n} \leq p(s_l^n)$, and that

$$\alpha_{ideal,s_l^n} \geq \frac{n_d+1}{2}g_{ideal} + \frac{1}{n_d}\sum_{i=1}^{n_d}(-p_{unif}), \tag{66}$$

it follows that $g_{ideal} \leq \alpha_{ideal,s_l^n} + p_{unif} \leq p(s_l^n) + p_{unif}$. Now using the Bhatia-Davis inequality in a similar way as for the proof of Theorem 10 in Section 11.4, we obtain the statistical inequality

$$Pr\Big(\Big|p(s_l^n) - \binom{m}{n}^{-1}\sum_{s_l^n \in \boldsymbol{S}} p(s_l^n)\Big| \leq \epsilon_{bias,s_l^n}\Big) \geq 1 - \frac{p_{unif}}{\epsilon^2_{bias,s_l^n}}.$$

This bound on the distance of $p(s_l^n)$ from p_{unif} can be used to upper bound g_{ideal}, as $g_{ideal} \leq p(s_l^n) + p_{unif}$ so then

$$Pr\Big(g_{ideal} \leq 2p_{unif} + \epsilon_{bias,s_l^n}\Big) \geq 1 - \frac{p_{unif}}{\epsilon^2_{bias,s_l^n}}.$$

Setting $\epsilon_{bias,s_l^n} = \binom{m}{n}^{-1/3}$ in the previous statistical inequality, it becomes

$$Pr\Big(g_{ideal} \leq 2p_{unif} + \binom{m}{n}^{-1/3}\Big) \geq 1 - \binom{m}{n}^{-1/3}.$$

Or in other words $g_{ideal} \in O(\binom{m}{n}^{-1/3})$ with exponentially high confidence. Now using Jensen's inequality,

$$\begin{aligned} D_0 &= \frac{1}{|\boldsymbol{S}|}\sum_{s_l^n \in \boldsymbol{S}} |p(s_l^n) - p_{unif}| \\ &\leq \left(\frac{1}{|\boldsymbol{S}|}\sum_{s_l^n \in \boldsymbol{S}} |p(s_l^n) - p_{unif}|^2\right)^{1/2} \\ &\leq ((1-p_{unif})p_{unif})^{1/2}, \end{aligned}$$

where the final line follows from applying the Bhatia-Davis inequality. From eqn. 67 we have that $g_{\text{avg}} \leq D_0$, and so by the previous inequality it follows that $g_{\text{avg}} \leq \binom{m}{n}^{-1/2}$. And so with exponentially high confidence $g_{ideal} \in O(\binom{m}{n}^{-1/3})$, and $g_{\text{avg}} \leq \binom{m}{n}^{-1/2}$.

The upper bound on the gradient bias error is then

$$\left|g_{ideal} - g_{\text{avg}}\right| \in O\left(\binom{m}{n}^{-1/3}\right),$$

with confidence $1 - O(m^{-n/3})$. □

We now derive the condition for when recycling mitigation with linear extrapolation can outperform postselection.

Theorem 26. *The condition:*

$$A\frac{n_d+1}{2}\left(\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4} + \binom{m}{n}^{-1/3}\right) + B\frac{n}{m-n+1}\sqrt{\frac{\binom{m}{n}}{n(1-\eta)^{n-1}\eta N_{tot}}} \leq C\sqrt{\frac{\binom{m}{n}}{(1-\eta)^n N_{tot}}},$$

with $A, B, C > 0$, defines a sampling regime where the sum of the worst-case statistical error and bias error of linear extrapolation recycling mitigation using the least squares method is less than the worst-case statistical error of postselection.

Proof. The linear extrapolation method consists of two iterations of linear least squares, with $n_d \in \{n_d \in \mathbb{Z}^+ | n_d < n\}$ data points used in both iterations. In the first iteration of least squares the data set $\{k, \tilde{D}_k\}_{k=1}^{n_d}$ is used to compute the gradient parameter g_{avg}. Where $\tilde{D}_k$ is the estimated absolute average deviation for the $n-k$–photon recycled distribution from statistics where k photons were lost. In the second iteration of least squares the data set $\{k, \tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^{n_d}$ is used to fit a linear model function that depends on g_{avg}, in order to generate the mitigated output for each s_l^n.

In the first iteration of least squares the linear model function used is

$$f(x_i, g_{\text{avg}}) = -\tilde{g}_{\text{avg}} x_i + \tilde{D}_0.$$

For which the residuals are of the form

$$r_i = y_i - (-\tilde{g}_{\text{avg}} x_i + \tilde{D}_0),$$

The sum of the squared residuals is a function of α_{s_n}, and may be written

$$\begin{aligned} S(\tilde{g}_{\text{avg}}) &= \sum_{i=1}^{n_d} r_i^2 \\ &= \sum_{i=1}^{n_d} (y_i - (-\tilde{g}_{\text{avg}} x_i + \tilde{D}_0))^2 \end{aligned}$$

The optimal solution may be found by taking the derivative of S with respect to $\tilde{g}_{\text{avg}}$

$$0 = \frac{\mathrm{d}S}{\mathrm{d}\tilde{g}_{\text{avg}}} = 2\sum_{i=1}^{n_d}(y_i - (-\tilde{g}_{\text{avg}}x_i + \tilde{D}_0))(x_i),$$

and then solving

$$\sum_{i=1}^{n_d}\tilde{g}_{\text{avg}}x_i^2 = \sum_{i=1}^{n_d}\tilde{D}_0 x_i - \sum_{i=1}^{n_d} y_i x_i.$$

Simplifying and substituting $\tilde{D}_i$ terms, this gives the optimal average gradient parameter as

$$\tilde{g}_{\text{avg}} = \frac{3}{2n_d+2}\tilde{D}_0 - \frac{6}{n_d(n_d+1)(2n_d+1)}\sum_{i=1}^{n_d}\tilde{D}_i x_i. \tag{67}$$

Using lemma 24, the statistical error of the average gradient parameter estimator, $\epsilon_{\text{hoeff},g}$, can be upper bounded

$$\begin{aligned}\left|\epsilon_{\text{hoeff},g}\right| &\leq |\tilde{g}_{\text{avg}} - g_{\text{avg}}| \\ &\in O\left(\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4}\right).\end{aligned}$$

The average gradient parameter estimator, $\tilde{g}_{\text{avg}}$, is used in a second iteration of least squares to generate the mitigated outputs. The data set used to generate the mitigated output for each s_l^n is defined as $\{x_i, y_i\}_{i=1}^{n_d} := \{k, \tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^{n_d}$. Each output string is assigned an individual model function of the form

$$f_{s_l^n}(x_i, \alpha_{s_l^n}) = \text{sgn}(-y_1)\tilde{g}_{\text{avg}}x_i + \alpha_{s_l^n}.$$

This function models the decay of the terms $\{\tilde{p}_R^k(s_l^n) - p_{unif}\}_{k=1}^{n_d}$ from the data set towards zero with increasing k. If $p_R^k(s_l^n) < p_{unif}$ then the gradient term (namely $\text{sgn}(-y_1)$) will be positive, and if $p_R^k(s_l^n) > p_{unif}$ the gradient will be negative. In the following analysis we will assume the sign, and therefore the gradient, is negative ($\text{sgn}(-y_1) = -1$) so that the model function is

$$f_{s_l^n}(x_i, \alpha_{s_l^n}) = -\tilde{g}_{\text{avg}}x_i + \alpha_{s_l^n},$$

noting that the analysis is the same for the case of positive sign. The optimal prefactor variable $\alpha_{s_l^n}$ to fit the model function to the data set is computed using the residuals

$$r_i = y_i - (-\tilde{g}_{\text{avg}}x_i + \alpha_{s_l^n}),$$

The sum of the squared residuals is a function of $\alpha_{s_l^n}$, and may be written

$$\begin{aligned}S(\alpha_{s_l^n}) &= \sum_{i=1}^{n_d} r_i^2 \\ &= \sum_{i=1}^{n_d}(y_i - (-\tilde{g}_{\text{avg}}x_i + \alpha_{s_l^n}))^2\end{aligned}$$

The value of $\alpha_{s_l^n}$ that minimises S may be found by taking the derivative of S with respect to $\alpha_{s_l^n}$, and solving

$$0 = \frac{\mathrm{d}S}{\mathrm{d}\alpha_{s_l^n}} = -2\sum_{i=1}^{n_d}(y_i - (-g_{\text{avg}}x_i + \alpha_{s_l^n}))$$

Which gives the optimal value as

$$\alpha_{s_l^n} = \frac{n_d+1}{2}\tilde{g}_{\text{avg}} + \frac{1}{n_d}\sum_{i=1}^{n_d} y_i. \tag{68}$$

And substituting in the data values this is then

$$\alpha_{s_l^n} = \frac{n_d+1}{2}\tilde{g}_{\text{avg}} + \frac{1}{n_d}\sum_{k=1}^{n_d}(\tilde{p}_R^k(s_l^n) - p_{unif}). \tag{69}$$

Now $\tilde{g}_{\text{avg}}$ can be decomposed into an ideal gradient parameter, g_{ideal}, the statistical error for the estimator, $\epsilon_{\text{hoeff},g}$, and the bias error from substituting the ideal gradient parameter for the average gradient parameter, $\epsilon_{\text{bias},g}$. The ideal gradient parameter, g_{ideal}, defines the gradient that results in the function outputting the ideal probability at $x_i = 0$ and $\alpha_{s_l^n} = p(s_l^n)$. And the recycled probability estimators $\tilde{p}_R^k(s_l^n)$ can be decomposed into recycled probabilities, $p_R^k(s_l^n)$, and their associated statistical errors, $\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}$. So that

$$\alpha_{s_l^n} = \frac{n_d+1}{2}\big(g_{ideal} + \epsilon_{\text{hoeff},g} + \epsilon_{\text{bias},g}\big) + \frac{1}{n_d}\sum_{i=1}^{n_d}(p_R^k(s_l^n) + \epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)} - p_{unif}). \tag{70}$$

While the ideal prefactor variable can be stated as

$$\alpha_{ideal,s_l^n} = \frac{n_d+1}{2}g_{ideal} + \frac{1}{n_d}\sum_{i=1}^{n_d}(p_R^k(s_l^n) - p_{unif}). \tag{71}$$

And the mitigated output is $p_{mit}(s_l^n) = p_{unif} + \alpha_{s_l^n}$. The magnitude of the error for the mitigated output is then

$$\begin{aligned}
|p_{mit}(s_l^n) - p(s_l^n)| &= |p_{unif} + \alpha_{s_l^n} - (p_{unif} + \alpha_{ideal,s_l^n})| \\
&= |\alpha_{s_l^n} - \alpha_{ideal,s_l^n}| \\
&\leq \frac{n_d+1}{2}(|\epsilon_{\text{hoeff},g}| + |\epsilon_{\text{bias},g}|) + |\epsilon_{\text{hoeff},\tilde{p}_R^1(s_l^n)}| \\
&\leq A\frac{n_d+1}{2}\left(\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4} + \binom{m}{n}^{-1/3}\right) + B\frac{n}{m-n+1}\sqrt{\frac{\binom{m}{n}}{n(1-\eta)^{n-1}\eta N_{tot}}}.
\end{aligned} \tag{72}$$

Where to get the second line a triangle inequality and that in the high loss regime

$$|\epsilon_{\text{hoeff},\tilde{p}_R^1(s_l^n)}| \geq \frac{1}{n_d}\sum_{i=1}^{n_d}|\epsilon_{\text{hoeff},\tilde{p}_R^k(s_l^n)}|$$

were used, and lemmas 24, 25 and 17, along with error term ϵ_{bias} that is bounded according to the inequality $|\epsilon_{bias}| \in O(\binom{m}{n}^{-1/3})$ derived in Appendix 25, were used to get the fourth line. And so the condition to beat postselection is

$$A\frac{n_d+1}{2}\left(\left(\frac{1}{(1-\eta)^n N_{tot}}\right)^{1/4} + \binom{m}{n}^{-1/3}\right) + B\frac{n}{m-n+1}\sqrt{\frac{\binom{m}{n}}{n(1-\eta)^{n-1}\eta N_{tot}}} \leq C\sqrt{\frac{\binom{m}{n}}{(1-\eta)^n N_{tot}}}.$$

□

We now provide a short proof of the following theorem from the main results section:

Theorem 5. *Assume that Conjectures 3 and 4 are true. Furthermore, assume that we are in the case where* $|p_{mit}(s) - p_{id}(s)| = \epsilon_{stat}(m,n,k,\eta,N_{tot}) + \kappa(n)\frac{1}{\binom{m}{n}}$, *with* $\kappa(n) < 1$, *and* $|p_{post}(s) - p_{id}(s)| = \sqrt{\frac{1}{(1-\eta)^n N_{tot}}}$, *where* $p_{mit}(s)$ *is the output probability of exponential extrapolation recycling mitigation. Then exponential extrapolation outperforms postselection for up to sample number* $N_{tot} \in O\Big(\frac{\binom{m}{n}^2}{\kappa(n)^2(1-\eta)^n}\Big)$ *and up to additive error* $\epsilon \in O\Big(\frac{\kappa(n)}{\binom{m}{n}}\Big)$.

Proof. Conjecture 4 being true implies that

$$\sqrt{\frac{1}{(1-\eta)^n N_{tot}}} - \epsilon_{stat}(m,n,k,\eta,N_{tot}) \geq c\sqrt{\frac{1}{(1-\eta)^n N_{tot}}},$$

for some $0 < c < 1$ for sufficiently large n. Thus, the condition that exponential extrapolation outperforms postselection, given by the following inequality holding $|p_{mit} - p_{id}(s)| \leq |p_{post}(s) - p_{id}(s)|$, is satisfied if $\kappa(n)\frac{1}{\binom{m}{n}} \leq c\sqrt{\frac{1}{(1-\eta)^n N_{tot}}}$. The bound on N_{tot} in the Theorem immediately follows. Furthermore, the bound on ϵ is directly obtained by replacing $(1-\eta)^n N_{tot} \in O\Big(\frac{\binom{m}{n}^2}{\kappa(n)^2}\Big)$ in $\frac{1}{\sqrt{(1-\eta)^n N_{tot}}}$, then using Stirling's approximation. □

14 Deterministic upper bound

14.1 Proof of Thm. 14

We now prove Thm. 14 from the main text:

For the class of unitary matrices U with submatrices A such that $p_{max} := \mathsf{max}_{s_l^n}(p(s_l^n)) = |\mathsf{Per}(A)|^2$, and where these matrices A satisfy $\frac{h_\infty^A}{\|A\|_2} \ll 1$, the bias error ϵ_{bias,s_l^n} is bounded

$$\epsilon_{bias,s_l^n} \in O(e^{-0.00002n}).$$

Proof. Let $p_{max} = \mathsf{max}_{s_l^n}(p(s_l^n))$, where $\max_{s_l^n}(.)$ denotes the maximum over the n–photon probability distribution. It is immediate to see that the interference term satisfies $I_{s_l^n,k} \leq p_{max}$. Furthermore, $p_{max} = |\mathsf{Per}(A)|^2$, where $A := (a_{ij})_{i,j\in\{1,\ldots,n\}}$ is a submatrix of the linear optical unitary U [67]. Note that $\|A\|_2 \leq \|U\|_2 \leq 1$, where $\|.\|_2$ denotes the spectral norm [72]. Let $h_\infty^A := \frac{1}{n}\sum_{i=1,\ldots,n}\|\mathbf{A}_i\|_\infty$, where $\mathbf{A}_i$ is the ith row of A, and $\|\mathbf{A_i}\|_\infty := \mathsf{max}_j(|a_{ij}|)$. Let $\|A\|_2 \leq T$. Using Thm. 2 in [62], we have that

$$|\mathsf{Per}(A)| \leq 2T^n e^{-0.00001(1-\frac{h_\infty^A}{\|A\|_2})^2 n}.$$

Since $\|A\|_2 \leq 1$, we immediately have

$$|\mathsf{Per}(A)| \leq 2e^{-0.00001(1-\frac{h_\infty^A}{\|A\|_2})^2 n}.$$

For classes of matrices where $\frac{h_\infty^A}{\|A\|_2} \ll 1$, we have that

$$|\mathsf{Per}(A)| \leq 2e^{-0.00001n},$$

and therefore that

$$p_{max} \leq 4e^{-0.00002n}.$$

Now, we have that $-p_{unif} \leq I_{s_l^n,k} - p_{unif} \leq p_{max} - p_{unif}$, and therefore the bias error is given by

$$\epsilon_1 = |I_{s_l^n,k} - p_{unif}| \leq \mathsf{max}\{|p_{max} - p_{unif}|, p_{unif}\}.$$

In the case where $|p_{max} - p_{unif}| < p_{unif}$, then we have an exponentially small bias error $\epsilon_{bias,s_l^n} \leq p_{unif} \leq \frac{1}{\binom{m}{n}} \leq e^{-Cnlogn}$, for some $C > 0$, when $m \in \Omega(n^{5.1})$ which is the boson sampling regime. Now, when $|p_{max} - p_{unif}| > p_{unif}$, we can use the bound for p_{max} we obtained, and show that

$$\epsilon_{bias,s_l^n} \leq |p_{max} - p_{unif}| \leq |4e^{-0.00002n} - e^{-Clog(n)}|,$$

and $|4e^{-0.00002n} - e^{-Clog(n)}| \in O(e^{-0.00002n})$. □

15 Richardson extrapolation methods for photon loss mitigation

In this section, we prove Thm. 16 as well as provide similar evidence (that the methods present no advantage over postselection) for different methods of performing extrapolation with increasing rates of photon loss.

First method of extrapolation at various noise rates

Let m be the number of modes of a linear optical circuit which can implement any $m \times m$ unitary. Into this circuit we input n photons in the first n modes. The notation $|n_1, \ldots, n_m\rangle$ denotes a state with n_i photons in the i^{th} mode, where $i \in \{1, \ldots, m\}$. $\eta \in]0, 1[$ is the probability to lose a photon in any given mode, and is the same for all modes. We want to compute a specific marginal probability $p(n_1 ... n_l|n)$ with $l \leq m$, and $\sum_{i=1,\ldots,l} n_i = c$, with $c \leq n$. The $|n$ indicates that we are computing the *ideal* marginal probability, when no photon is lost. Let $p(n_1 \ldots n_l \cap j)$ be the probability of observing the output $(n_1, \ldots, n_l)$ and detecting j photons in all m modes, with $j \in \{c, \ldots, n\}$. When postselecting on no photons being lost, we are computing

$$p(n_1 \ldots n_l \cap n) = (1-\eta)^n p(n_1 \ldots n_l|n).$$

However, if we compute $p(n_1 \ldots n_l)$ without caring about whether no photon is lost, we end up computing

$$p_\eta(n_1 ... n_l) = \sum_{i=0,\ldots,n-c} (1-\eta)^{n-i} \eta^i p(n_1 \ldots n_l|n-i). \tag{73}$$

Extrapolation techniques consist of estimating $p_\eta(n_1 \ldots n_l)$ for different values of η, then deducing from these an estimate of $p(n_1 \ldots n_l|n)$. One example of how this can be done is the Richardson extrapolation technique, at the heart of the zero noise extrapolation (ZNE) approach. Rather interestingly, we will show that these techniques offer no advantage over post-selection in terms of estimating $p(n_1 \ldots n_l|n)$. Let $\alpha_i := p(n_1 \ldots n_l|n-i)$, we can then write

$$p_\eta(n_1 \ldots n_l) := \sum_{i=0,\ldots,n-c} (1-\eta)^{n-i} \eta^i \alpha_i. \tag{74}$$

A natural way to estimate α_0 through extrapolation would be to compute an estimate $\tilde{p}_\eta(n_1 \ldots n_l)$ of $p_\eta(n_1 \ldots n_l)$ for $n-c+1$ values η_i of η, with $i \in \{0, \ldots, n-c\}$, and (by convention) $\eta_{i+1} > \eta_i$. We will deal with additive error estimates, that is, $\tilde{p}_\eta(n_1 \ldots n_l) = p_\eta(n_1 \ldots n_l) + \epsilon$, with $|\epsilon| \leq \epsilon_{max}$, and $\epsilon_{max} \in [0, 1]$ is the additive error estimate. Note that an ϵ_{max} additive error estimate $\tilde{p}_\eta(n_1 \ldots n_l)$ can be obtained with high probability from $O(\frac{1}{\epsilon_{max}^2})$ runs of the boson sampling device, by Hoeffding's inequality [56].

Performing the above mentioned extrapolation strategy, we obtain the following set of equations, written in matrix form, to be solved for obtaining an estimate of α_0

$$\begin{pmatrix} (1-\eta_0)^n & \eta_0(1-\eta_0)^{n-1} & \ldots & \eta_0^{n-c}(1-\eta_0)^k \\ . & . & \ldots & . \\ . & . & \ldots & . \\ (1-\eta_{n-c})^n & \eta_{n-c}(1-\eta_{n-c})^{n-1} & \ldots & \eta_{n-c}^{n-c}(1-\eta_{n-c})^c \end{pmatrix} \begin{pmatrix} \tilde{\alpha}_0 \\ . \\ . \\ \tilde{\alpha}_{n-c} \end{pmatrix} = \begin{pmatrix} p_{\eta_0}(n_1 \ldots n_l) \\ . \\ . \\ p_{\eta_{n-c}}(n_1 \ldots n_l) \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ . \\ . \\ \epsilon_{n-c} \end{pmatrix}, \tag{75}$$

with $|\epsilon_i| \leq \epsilon_{max}$, and $\tilde{\alpha}_i$ is an estimate of α_i obtained from using the estimates $\tilde{p}_{\eta_i}(n_1 \ldots n_l) = p_{\eta_i}(n_1 \ldots n_l) + \epsilon_i$ Let,

$$L := \begin{pmatrix} (1-\eta_0)^n & \eta_0(1-\eta_0)^{n-1} & \ldots & \eta_0^{n-c}(1-\eta_0)^c \\ . & . & \ldots & . \\ . & . & \ldots & . \\ (1-\eta_{n-c})^n & \eta_{n-c}(1-\eta_{n-c})^{n-1} & \ldots & \eta_{n-c}^{n-c}(1-\eta_{n-k})^c \end{pmatrix}.$$

We can rewrite L as

$$L = DW, \tag{76}$$

with

$$D = \begin{pmatrix} (1-\eta_0)^n & 0 & 0 & \dots 0 \\ 0 & (1-\eta_1)^n & 0 & \dots 0 \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & 0 & \dots (1-\eta_{n-c})^n \end{pmatrix} \tag{77}$$

a diagonal matrix, and

$$W = \begin{pmatrix} 1 & \frac{\eta_0}{1-\eta_0} & \dots & (\frac{\eta_0}{1-\eta_0})^{n-c} \\ 1 & \frac{\eta_1}{1-\eta_1} & \dots & (\frac{\eta_1}{1-\eta_1})^{n-c} \\ . & . & . & . \\ . & . & . & . \\ 1 & \frac{\eta_{n-c}}{1-\eta_{n-c}} & \dots & (\frac{\eta_{n-c}}{1-\eta_{n-c}})^{n-c} \end{pmatrix} \tag{78}$$

a Vandermonde matrix [39]. Multiplying both sides of eqn. 75 by L^{-1} and invoking standard matrix multiplication rules, we obtain

$$\tilde{\alpha}_0 = \sum_{i=0,n-c} L^{-1}_{1i} p_{\eta_i}(n_1 \dots n_l) + \sum_{i=0,\dots,n-c} L^{-1}_{1i} \epsilon_i, \tag{79}$$

where L^{-1}_{1i} is the element of the first row and ith column of L^{-1}. Note that, by construction of our method, $\eta_i \neq \eta_{i+1}$, then D is invertible, and so is W [39], thus L^{-1} always exists. Furthermore, $\alpha_0 = \sum_{i=0,n-k} L^{-1}_{1i} p_{\eta_i}(n_1 \dots n_l)$, since exactly computing the probabilities $p_{\eta_i}(n_1 \dots n_l)$ will lead to an exact computation of α_0. Therefore, the error associated to our extrapolation technique is given by

$$E_{extrap} := |\sum_{i=0,\dots,n-k} L^{-1}_{1i} \epsilon_i|. \tag{80}$$

Note that the overall sample complexity of the extrapolation protocol is $O(\frac{n-c+1}{\epsilon^2_{max}})$. For post-selection, an ϵ_{max} additive error estimate $\tilde{p}_\eta(n_1 \dots n_l \cap n)$ requires $O(\frac{1}{\epsilon^2_{max}})$ samples, and can be performed for $\eta = \eta_0$, that is without artificially increasing loss. From eqn. 73, we see that post-selection induces an error of

$$E_{post} := |\frac{\epsilon}{\sqrt{(1-\eta_0)^n}}| \leq \frac{\epsilon_{max}}{\sqrt{(1-\eta_0)^n}}, \tag{81}$$

with $|\epsilon| \leq \epsilon_{max}$.

In the remainder of this section, we will compute an upper bound for E_{extrap}, and show that this upper bound is greater than the corresponding upper bound for E_{post} shown in eqn. 81. This gives strong evidence that the error induced by extrapolation is higher than that of postselection for a comparable number of samples, and therefore that extrapolation offers no advantage over postselection. Although the upper bound argument we show gives strong evidence that extrapolation techniques are not advantageous when compared to postselection, we will provide further evidence that this is the case. In particular, for a random distribution of errors $\{\epsilon_i\}$ with $|\epsilon_i| \leq \epsilon_{max}$, we numerically show that the condition

$$E_{extrap} > \frac{\epsilon_{max}}{(1-\eta_0)^n} \tag{82}$$

is never violated after some value of n, confirming our analytical results. We start by the analytical upper bound argument. By a triangle inequality,

$$E_{extrap} \leq ||L^{-1}||_\infty \epsilon_{max},$$

with $||L^{-1}||_\infty := \mathsf{max}_i \sum_j |L^{-1}_{ij}|$. Also,

$$||L^{-1}||_\infty \leq ||D^{-1}||_\infty ||W^{-1}||_\infty.$$

From the definition of $||.||_\infty$, we can directly show

$$||D^{-1}||_\infty = \frac{1}{(1-\eta_{n-c})^n}.$$

For $||W^{-1}||_\infty$, we can use an upper bound on the norm of Vandermonde matrices shown in [39]

$$||W^{-1}||_\infty \leq \mathsf{max}_i \prod_{j\neq i=0,\ldots,n-c} \frac{1+\frac{\eta_i}{1-\eta_i}}{|\frac{\eta_i}{1-\eta_i}-\frac{\eta_j}{1-\eta_j}|}.$$

Thus,

$$||L^{-1}||_\infty \leq \frac{1}{(1-\eta_{n-c})^n}\mathsf{max}_i \prod_{j\neq i=0,\ldots,n-c} \frac{1+\frac{\eta_i}{1-\eta_i}}{|\frac{\eta_i}{1-\eta_i}-\frac{\eta_j}{1-\eta_j}|}.$$

and

$$E_{extrap} \leq \epsilon_{max}\frac{1}{(1-\eta_{n-c})^n}\mathsf{max}_i \prod_{j\neq i=0,\ldots,n-c} \frac{1+\frac{\eta_i}{1-\eta_i}}{|\frac{\eta_i}{1-\eta_i}-\frac{\eta_j}{1-\eta_j}|}.$$

We will now show that the upper bound on E_{extrap} is greater than that of E_{post}. This is quantified through the following theorem .

Theorem 27. *For $\eta_0,\ldots,\eta_{n-c}$ with $\eta_{i+1}>\eta_i$ and $\eta_0\geq 0$, $\eta_{n-c}<1$ the following holds*

$$\frac{1}{(1-\eta_{n-c})^n}\mathsf{max}_i \prod_{j\neq i=0,\ldots,n-c} \frac{1+\frac{\eta_i}{1-\eta_i}}{|\frac{\eta_i}{1-\eta_i}-\frac{\eta_j}{1-\eta_j}|} \geq \frac{1}{(1-\eta_0)^n}. \tag{83}$$

Proof. First, note that $\frac{\eta}{1-\eta}$ is a monotonically increasing function of η, thus $\frac{\eta_i}{1-\eta_i}<\frac{\eta_{i+1}}{1-\eta_{i+1}}$. This allows us to lower bound $\mathsf{max}_i\prod_{j\neq i=0,\ldots,n-k}\frac{1+\frac{\eta_i}{1-\eta_i}}{|\frac{\eta_i}{1-\eta_i}-\frac{\eta_j}{1-\eta_j}|}$ as

$$\mathsf{max}_i \prod_{j\neq i=0,\ldots,n-c} \frac{1+\frac{\eta_i}{1-\eta_i}}{|\frac{\eta_i}{1-\eta_i}-\frac{\eta_j}{1-\eta_j}|} \geq \frac{(1+\frac{\eta_0}{1-\eta_0})^{n-c}}{(\frac{\eta_{n-c}}{1-\eta_{n-c}}-\frac{\eta_0}{1-\eta_0})^{n-c}}.$$

Our strategy is to show that the following holds

$$\frac{1}{(1-\eta_{n-c})^n}\frac{(1+\frac{\eta_0}{1-\eta_0})^{n-c}}{(\frac{\eta_{n-c}}{1-\eta_{n-c}}-\frac{\eta_0}{1-\eta_0})^{n-c}} \geq \frac{1}{(1-\eta_0)^n} \tag{84}$$

eqn. 84 being true implies that eqn. 83 is also true, and thus is sufficient for proving Thm. 27. eqn. 84 can be rewritten as

$$\frac{(1-\eta_0)^n}{(1-\eta_{n-c})^n}\frac{(1-\eta_0)^{n-c}}{(\frac{\eta_{n-c}}{1-\eta_{n-c}}-\frac{\eta_0}{1-\eta_0})^{n-c}} \geq 1. \tag{85}$$

rewriting the left hand side of the above eqn.

$$\frac{(1-\eta_0)^n}{(1-\eta_{n-c})^n}\frac{(1+\frac{\eta_0}{1-\eta_0})^{n-c}}{(\frac{\eta_{n-c}}{1-\eta_{n-c}}-\frac{\eta_0}{1-\eta_0})^{n-c}} = \frac{(1-\eta_0)^c}{(1-\eta_{n-c})^c}\Big(\frac{(1-\eta_0)(1+\frac{\eta_0}{1-\eta_0})}{(1-\eta_{n-c})(\frac{\eta_{n-c}}{1-\eta_{n-c}}-\frac{\eta_0}{1-\eta_0})}\Big)^{n-c},$$

and observing that

$$\frac{(1-\eta_0)(1+\frac{\eta_0}{1-\eta_0})}{(1-\eta_{n-c})(\frac{\eta_{n-c}}{1-\eta_{n-c}}-\frac{\eta_0}{1-\eta_0})} = \frac{1-\eta_0}{\eta_{n-c}-\eta_0},$$

and plugging this into eqn. (85) we obtain

$$\frac{(1-\eta_0)^c}{(1-\eta_{n-c})^c}\Big(\frac{1-\eta_0}{\eta_{n-c}-\eta_0}\Big)^{n-c} \geq 1. \tag{86}$$

Now, $1-\eta_0\geq 1-\eta_{n-c}$ and $1-\eta_0\geq\eta_{n-c}-\eta_0$ since $\eta_0<\eta_{n-c}<1$. This implies that eqn. 86 is true, and thus eqn. s 85 and 84 hold, and therefore Thm. 27 is proved. □

Second method of extrapolation at various noise rates

Another possible extrapolation technique can be performed by considering eqn. (73) where the $(1-\eta_i)^{n-i}$ terms are expanded in order to obtain

$$p_\eta(n_1 \dots n_l) = \sum_{i=0,\dots,n} \beta_i \eta^i. \tag{87}$$

Where $\beta_0 = \alpha_0$, and β_is are linear combinations of the α_is defined previously. The extrapolation procedure proceeds in a similar manner to that described above, but now we compute $p_\eta(n_1 \dots n_l)$ for $n+1$ values of loss $\eta_n > \eta_{n-1} \cdots > \eta_0$ to solve for the coefficients $\{\beta_i\}$, with the matrix L in this case given by

$$L = \begin{pmatrix} 1 & \eta_0 & \dots & \eta_0^n \\ 1 & \eta_1 & \dots & \eta_1^n \\ . & . & . & . \\ . & . & . & . \\ 1 & \eta_n & \dots & \eta_n^n \end{pmatrix}. \tag{88}$$

L is a Vandermonde matrix, and we can directly use the result of [39] to upper bound $||L^{-1}||_\infty$ as

$$||L^{-1}||_\infty \leq \mathsf{max}_i \prod_{j \neq i=0,\dots,n} \frac{1+\eta_i}{|\eta_i - \eta_j|}.$$

Therefore,

$$E_{extrap} \leq \epsilon_{max} \mathsf{max}_i \prod_{j \neq i=0,\dots,n} \frac{1+\eta_i}{|\eta_i - \eta_j|}$$

We will now prove that the upper bound on E_{extrap} is larger than that of E_{post} for this method of extrapolation, giving strong evidence that this technique offers no advantage over post-selection. This amounts to proving the following.

Theorem 28. *For $\eta_0, \dots, \eta_n$ with $\eta_{i+1} > \eta_i$ and $\eta_0 \geq 0$, $\eta_n < 1$ the following holds*

$$\mathsf{max}_i \prod_{j \neq i=0,\dots,n} \frac{1+\eta_i}{|\eta_i - \eta_j|} \geq \frac{1}{(1-\eta_0)^n}. \tag{89}$$

Proof. We note that

$$\mathsf{max}_i \prod_{j \neq i=0,\dots,n} \frac{1+\eta_i}{|\eta_i - \eta_j|} \geq \frac{(1+\eta_0)^n}{(\eta_n - \eta_0)^n},$$

and that $\frac{(1+\eta_0)^n}{(\eta_n-\eta_0)^n} \geq \frac{1}{(1-\eta_0)^n}$ since $\frac{1}{\eta_n-\eta_0} > \frac{1}{1-\eta_0}$ and $1+\eta_0 > 1$. This completes the proof. □

For this technique as well we numerically compute the number of violations of $E_{extrap} > \frac{\epsilon_{max}}{(1-\eta_0)^n}$ and plot these in Figure 7. We took $\epsilon_{max} = 0.01$, $\eta_0 = 0.01$, $\eta_n = 0.95$, and η_i for $0 < i < n$ equally spaced. We varied the value of n between 3 and 16. For each value of n we performed 3000 runs, where at each run we took $n+1$ values of $\{\epsilon_i\}$ chosen uniformly randomly from $[-\epsilon_{max}, \epsilon_{max}]$. As can be observed in Figure 7, the number of violations approaches zero with increasing n, confirming that this extrapolation performs worse than post-selection after some value of n. A similar behaviour is observed for different values of ϵ_{max}, η_0 and η_n.

Proof of Thm. 16

We now prove Thm. 16 from the main text:

For all $n \geq n_0$, with n_0 a positive integer, $\mathsf{M}(E_{extrap}) \geq \frac{\epsilon_{max}}{\sqrt{(1-\eta)^n}}$.

A third extrapolation technique which can be used is the Richardson extrapolation technique, at the heart of the zero noise extrapolation method for mitigating qubit errors, which has gained widespread use and is at the heart of ZNE [36]. This method uses the expansion of eqn. (87) with

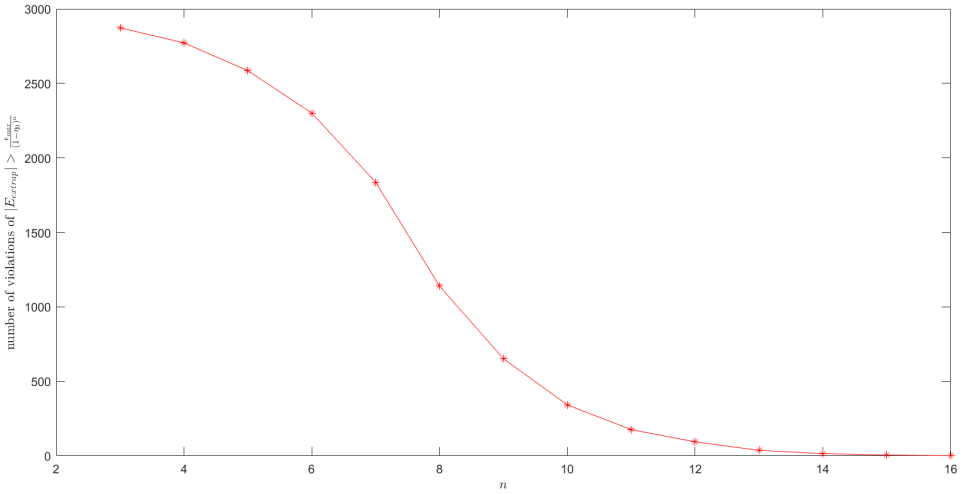


Figure 7: Number of violations of $|E_{extrap}| \geq \frac{\epsilon_{max}}{(1-\eta_0)^n}$ plotted versus n (see main text).

$\eta_i = c_i\eta$, for $i \in \{0, \ldots, n\}$ with $\eta \in [0, 1]$, $c_0 = 1$, and c_i are positive reals satisfying $c_{i+1} > c_i$. The Richardson extrapolation method consists of computing an estimate $\tilde{p}(n_1 \ldots n_l|n)$ of $p(n_1 \ldots n_l|n)$ as follows

$$\tilde{p}(n_1 \ldots n_l|n) = \sum_{i=0\ldots n} \gamma_i \tilde{p}_{\eta_i}(n_1 \ldots n_l),$$

where, $\gamma_i := (-1)^n \prod_{j\neq i} \frac{c_j}{c_i - c_j}$. It can be shown that [13] $\sum_{i=0,\ldots,n} \gamma_i = 1$ and $\sum_{j=0,\ldots,n} \gamma_i c_i^j = 0$ for $j = 1, \ldots, n$. Furthermore, it can also be shown that $p(n_1 \ldots n_l|n) = \sum_{i=0,\ldots n} \gamma_i p_{\eta_i}(n_1 \ldots n_l)$. Thus, the error associated with this technique is given by

$$E_{extrap} := |\sum_{j=0,\ldots n} \gamma_i \epsilon_i|.$$

In the remainder of this paragraph, we will show that $\gamma_{i-1} = L_{1i}^{-1}$ for $i \in \{1, \ldots, n+1\}$ where L is the Vandermonde matrix of eqn. (88) with $\eta_i = c_i\eta$. This means that the results of the previous section follow through, and therefore that Richardson extrapolation offers no advantage over post-selection.

Let $\mathbf{B} := \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_n \end{pmatrix}$, and $\mathbf{P} = \begin{pmatrix} p_{\eta_0}(n_1 \ldots n_l) \\ p_{\eta_1}(n_1 \ldots n_l) \\ . \\ . \\ . \\ p_{\eta_n}(n_1 \ldots n_l) \end{pmatrix}$. In the case of a perfect computation of the probabilities $p_{\eta_i}(n_1 \ldots n_l)$ (i.e $\epsilon_i = 0$), the extrapolation technique based on the matrix L of eqn. (88) amounts to computing β_0 from the following system of eqn. s

$$\mathbf{B} = L^{-1}\mathbf{P}.$$

Invoking standard multiplication rules for the above equation, one obtains

$$p(n_1 \ldots n_l) = \beta_0 = \sum_{i=1,\ldots n+1} L_{1i}^{-1} p_{\eta_{i-1}}(n_1 \ldots n_l).$$

Plugging the expansion $p_{\eta_j}(n_1 \ldots n_l) = \sum_{i=0,\ldots n} \beta_i \eta_i^j$ and $\eta_i = c_i\eta$ into the above equation, one obtains

$$\beta_0 = \sum_{i=1,\ldots n+1} L_{1i}^{-1}\beta_0 + \sum_{j=1\ldots,n} \sum_{i=1\ldots n+1} L_{1i}^{-1} c_{i-1}^j \beta_j \eta^j.$$

By a direct identification of the left hand side of the above equation with its right hand side, we find that $\sum_{i=1\ldots n+1} L_{1i}^{-1} = 1$, and $\sum_{i=1,\ldots,n+1} L_{1i}^{-1} c_{i-1}^{j} = 0$ for $j \in \{1 \ldots n\}$. These sets of equations are exactly those defining the coefficients γ_i (as seen previously) and allowing to uniquely determine them, thus we can make the identification $L_{1i}^{-1} = \gamma_{i-1}$ for $i \in \{1, \ldots, n\}$ and our result is demonstrated.

16 Lyapunov bound

16.1 Prospects of improving bounds on the interference terms

A natural question is whether one can tighten the bounds, beyond what is guaranteed from Thm. 8, on the error term ϵ_{bias,s_l^n} incurred by replacing the interference term $I_{s_l^n,k}$ with its average over the Haar measure.

In this section, we explore one attempt to tighten the bound on ϵ_{bias,s_l^n}. We will work in the no-collision regime where $m \gg n^2$, so that we can approximate output probabilities as moduli squared of permanents of i.i.d. Gaussian matrices, with appropriate rescaling. In this regime, the interference term $I_{s_l^n,k}$ can be thought of as a sum of N'_k random variables

$$X_i := \frac{|\mathsf{Per}(G_i)|^2}{m^n}, \tag{90}$$

for $i \in \{1, \ldots, N'_k\}$, where G_i is an i.i.d. Gaussian $n \times n$ matrix with entries chosen independently from $\mathcal{N}_{\mathbb{C}}(0, 1)$. Note that for some i and j, it is possible that G_i or G_j share some rows in common, or are even equal. This corresponds to the fact that the interference term is in general a sum of probabilities of output bit strings sharing some overlap (meaning their associated permanents have common rows [6]). This means that the random variables X_i need not all be independent. Nevertheless, we will look at a sum of independent random variables X_i having the form of eqn. (90). Our reason for working with an independent sum is that it simplifies the analysis, while giving an intuition about what to expect in the more general case of possibly dependent X_i's. We will further comment on this point below.

In the remainder of this section, we will show that a $\mathsf{poly}(n)$ sized sum of independent X_i's distributed according to eqn. (90) *does not* verify a sufficient condition, the Lyapunov condition [73], for the central limit theorem (CLT) to hold as $n \to \infty$. Our proof relies on a conjecture of [55] on the expectation value $\mathbf{E}(X_i^t)$ over the set of Gaussian matrices G_i, where $t \in \mathbb{N}$ and $t > 2$. We also provide numerical evidence that the distribution of this sum is indeed not a normal distribution. Our result shows that it is non-trivial to improve the bound on ϵ_{bias,s_l^n} by trying to link $I_{s_l^n,k}$ to some known probability distribution, which was our initial motivation for trying to prove a CLT convergence result.

Let Y_i, $i \in \{1, \ldots, N\}$ be real, independent and identically distributed, random variables satisfying $\mathbf{E}(Y_i) = 0$, where $\mathbf{E}(.)$ denotes expectation value over the distribution of the Y_i's. Let

$$S_N := \sum_{i=1,\ldots,N} Y_i, \tag{91}$$

and

$$\sigma_N := \sqrt{\mathbf{E}(S_N^2)}. \tag{92}$$

Furthermore, for any $r > 2$ and $r \in \mathbb{N}$ let

$$\mathbf{E}(|Y|^r) := \mathbf{E}(|Y_i|^r), \tag{93}$$

for all $i \in \{1, \ldots, N\}$. The Lyapunov condition can be stated in this case as [73]

Theorem 29. (Lyapunov condition) *If for some fixed $r > 2$,*

$$N \frac{\mathbf{E}(|Y|^r)}{\sigma_N^r} \to 0, \tag{94}$$

then as $N \to \infty$

$$\frac{S_N}{\sigma_N} \to^d \mathcal{N}(0,1). \tag{95}$$

Where $\to^d$ *denotes convergence of the distribution of* $\frac{S_N}{\sigma_N}$.

For dependent identically distributed random variables, which correspond to the probabilities constituting the interference term, the Lyapunov condition becomes stricter to verify. In particular, for a specific type of dependence, $M(n)$-dependent random variables [73], the numerator in eqn. (94) is multiplied by $M(n)^{r-1}$, where $M(n) > 1$ and $M(n) \in \mathbb{N}$ is an integer whose value can depend on n [73]. One would expect, as mentioned earlier, that the non-convergence results established here for the case of independent random variables hold as well for the dependent case, although we do not formally prove this.

A final ingredient we will use is the following conjecture appearing in [55] (Section 4.8, Conjecture 4.12), and whose truth is supported by numerical simulations performed in [55].

Conjecture 30. *For* $n, t > 2$, $n, t \in \mathbb{N}$

$$\mathbf{E}_{G \in \mathcal{G}_{n\times n}}(|\mathsf{Per}(G)|^{2t}) \in \Theta\left(\frac{(n!)^{2t}(t!)^{2n}}{(nt)!}\right). \tag{96}$$

Where $\mathrm{E}_{G\in\mathcal{G}_{n\times n}}(.)$ is the expectation value over the set of $n \times n$ Gaussian matrices with entries from $\mathcal{N}_{\mathbb{C}}(0,1)$. For simplicity, we will henceforth denote $\mathbf{E}_{G\in\mathcal{G}_{n\times n}}(.)$ as $\mathbf{E}(.)$. Let

$$Y_i := X_i - \frac{n!}{m^n}, \tag{97}$$

where X_i are as defined in eqn. (90), and are independent, identically distributed random variables, $i \in \{1, \ldots, N\}$. It is immediate to observe that $\mathbf{E}(Y_i) = 0$. We show the following.

Theorem 31. *For all* $r > 2$, $n \gg 1$, *and for the independent, identically distributed random variables* Y_i *defined in eqn. (97), we have that*

$$N\frac{\mathbf{E}(|Y|^r)}{\sigma_N^r} \geq A\frac{\beta(r)^n}{N^{\frac{r}{2}-1}}, \tag{98}$$

where $\beta(r) > 1$ *is a positive real number dependent on* r, *and* $A > 0$ *a constant.*

When $N = \mathsf{poly}(n)$, Thm. 31 shows that Lyapunov condition is not satisfied. Although this condition is sufficient, but not necessary, for the CLT to hold we provide numerical evidence that $N = n^2$, $N = n^3$, and $N = n^4$ sized sums of i.i.d. Gaussian matrices do not converge to a normal distribution for $n \in \{2, 3, 4, 5\}$. We plot our results for $N = n^3$ and $n \in \{2, 3, 4, 5\}$ in Figures 8 (a)-(d).

It is interesting to note that in Thm. 31, when $N \in O(\mathsf{exp}(n))$, the lower bound on the Lyapunov condition can converge to 0 as $n \to \infty$. Marginal probabilities corresponding to a large number of lost photons are sums of, possibly exponential, numbers of probabilities having the form of eqn. (90). Our result provides evidence that these marginals are asymptotically normally distributed, and therefore efficient to sample from. Although low order as well as high loss marginals of boson sampling are known to be easy to compute and sample from [58, 74, 61], our result might provide a new perspective on simulating lossy boson sampling marginals by linking these to normally distributed random variables.

16.2 Proof of Thm. 31

We now prove Thm. 31 .

Proof. Fix an $r > 2$. We begin by evaluating σ_N^r.

$$\sigma_N^2 = (\mathbf{E}(S_N^2)) = \mathbf{E}((\sum_{i=1,\ldots,N} Y_i)^2) = \sum_{i=1,\ldots,N} \mathbf{E}(Y_i^2) + 2\sum_{i=1,\ldots,N}\sum_{j>i} \mathbf{E}(Y_iY_j).$$

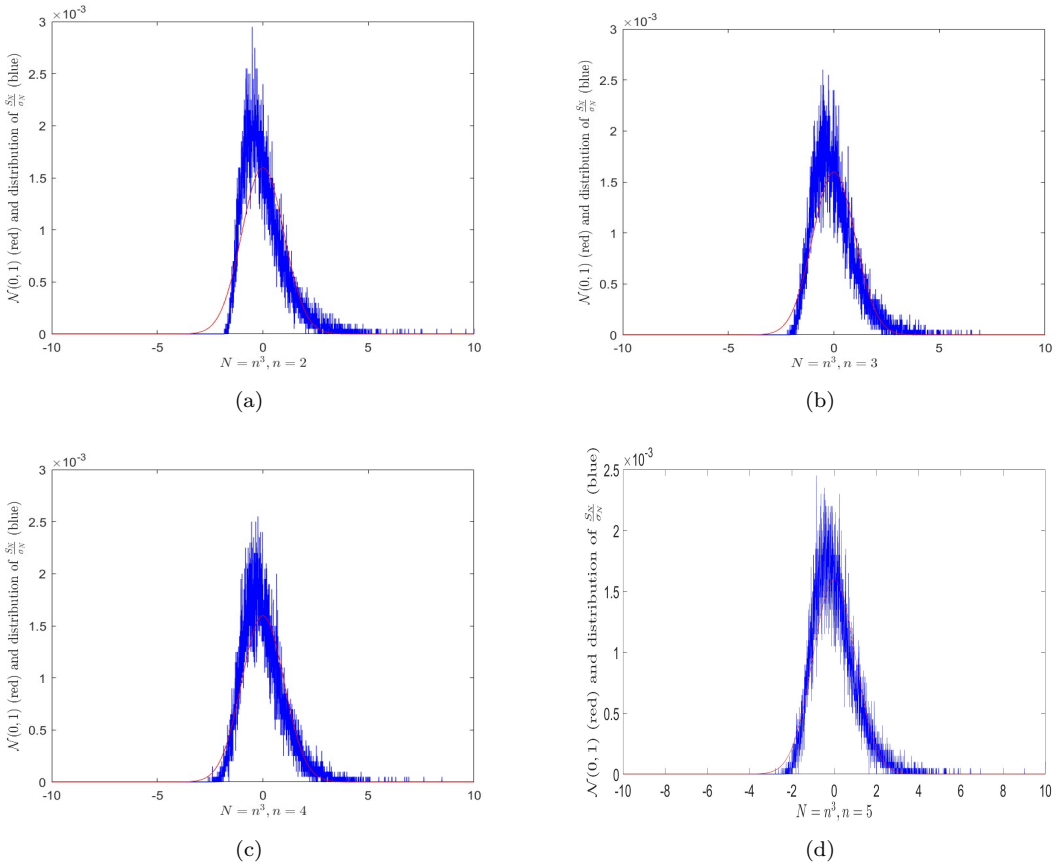


Figure 8: (a)-(b) *Comparison of distributions of* $\frac{S_N}{\sigma_N}$ *and the normal* $\mathcal{N}(0,1)$ *distribution.* (a) Distribution of $\frac{S_N}{\sigma_N}$ compared to the normal $\mathcal{N}(0,1)$ distribution. $n = 2$, $N = n^3$, and 20000 samples of i.i.d. Gaussian matrices were used to construct the distribution of $\frac{S_N}{\sigma_N}$. (b) Distribution of $\frac{S_N}{\sigma_N}$ compared to the normal $\mathcal{N}(0,1)$ distribution. $n = 3$, $N = n^3$, and 20000 samples of i.i.d. Gaussian matrices were used to construct the distribution of $\frac{S_N}{\sigma_N}$. (c) Distribution of $\frac{S_N}{\sigma_N}$ compared to the normal $\mathcal{N}(0,1)$ distribution. $n = 4$, $N = n^3$, and 20000 samples of i.i.d. Gaussian matrices were used to construct the distribution of $\frac{S_N}{\sigma_N}$. (d) Distribution of $\frac{S_N}{\sigma_N}$ compared to the normal $\mathcal{N}(0,1)$ distribution. $n = 5$, $N = n^3$, and 20000 samples of i.i.d. Gaussian matrices were used to construct the distribution of $\frac{S_N}{\sigma_N}$.

The $Y_i's$ are independent identically distributed, and $\mathbf{E}(Y_i) = 0$ by construction, therefore $\mathbf{E}(Y_iY_j) = \mathbf{E}(Y_i)\mathbf{E}(Y_j) = 0$, $\forall i \neq j$, and $\sum_{i=1,\dots,N} \mathbf{E}(Y_i^2) = N\mathbf{E}(Y^2) = N\mathbf{E}(X - \frac{n!}{m^n})^2 = N\big(\mathbf{E}(X^2) - 2\mathbf{E}(X)\frac{n!}{m^n} + (\frac{n!}{m^n})^2\big)$. Now, $\mathbf{E}(X^2) = \frac{(n!)^2(n+1)}{m^{2n}}$ and $\mathbf{E}(X) = \frac{n!}{m^n}$ [55]. Thus

$$\sigma_N^2 = N(\frac{n(n!)^2}{m^{2n}}),$$

and then

$$\sigma_N^r = (N(\frac{n(n!)^2}{m^{2n}}))^{\frac{r}{2}}.$$

We will now compute a lower bound $\mathbf{E}(|Y|^r)$. First, we will use a triangle inequality,

$$\mathbf{E}(|Y|^r) \geq |\mathbf{E}(Y^r)|.$$

We now focus on computing $|\mathbf{E}(Y^r)|$. Let $\mu := \frac{n!}{m^n}$

$$\mathbf{E}(Y^r) = \mathbf{E}(X-\mu)^r = \sum_{i=0,\dots,r} \binom{r}{i} \mathbf{E}(X^i)\mu^{r-i}(-1)^{r-i} = \sum_{i=0,\dots,r} \binom{r}{i} \frac{\mathbf{E}(|\mathsf{Per}(G)|^{2i})\mu^{r-i}(-1)^{r-i}}{m^{ni}}.$$

Plugging conjecture 30 into $\mathbf{E}(Y^r)$, we obtain

$$\mathbf{E}(Y^r) = \sum_{i=0,\dots,r} \binom{r}{i} \frac{A\frac{(n!)^{2i}(i!)^{2n}}{(ni)!}(\frac{n!}{m^n})^{r-i}(-1)^{r-i}}{m^{ni}} = \frac{1}{m^{nr}} \sum_{i=0,\dots,r} (-1)^{r-i} B_i \frac{\binom{r}{i}(n!)^{r+i}(i!)^{2n}}{(ni)!},$$

for some constants $A, B_i > 0$ for $i \in \{0, \dots, r\}$. Since $n \gg 1$, we can use a Stirling approximation for $i > 0$

$$\frac{(n!)^{r+i}}{(ni)!} \approx \frac{(\frac{n}{e})^{nr+ni}(\sqrt{2\pi n})^{r+i}}{(\frac{ni}{e})^{ni}\sqrt{2\pi ni}} = \left(\frac{1}{i}\right)^{ni}\left(\frac{n}{e}\right)^{nr}\frac{(2\pi n)^{\frac{i+r}{2}}}{\sqrt{2\pi ni}}.$$

Plugging this into $\mathbf{E}(Y^r)$, we get

$$\mathbf{E}(Y^r) = \frac{(\frac{n}{e})^{nr}}{m^{nr}}\left((-1)^r A(2\pi n)^{\frac{r}{2}} + \sum_{i=1,\dots,r}(-1)^{r-i}B_i\binom{r}{i}(i!)^{2n}\frac{1}{i^{ni}}\frac{(2\pi n)^{\frac{i+r}{2}}}{\sqrt{2\pi ni}}\right).$$

For any $i \leq r$, $\frac{(2\pi n)^{\frac{i+r}{2}}}{\sqrt{2\pi ni}} \in O(n^r)$, we can simplify the above to

$$\mathbf{E}(Y^r) = A\frac{n^r(\frac{n}{e})^{nr}}{m^{nr}}\left(B(-1)^r + \sum_{i=1,\dots,r}(-1)^{r-i}C_i\binom{r}{i}(i!)^{2n}\frac{1}{i^{ni}}\right),$$

for some constants $A, B, C_i > 0$. Now, we will look at $|B(-1)^r + \sum_{i=1,\dots,r}(-1)^{r-i}C_i\binom{r}{i}(i!)^{2n}\frac{1}{i^{ni}}))| = |(-1)^rB + \sum_{i=1,\dots,r}(-1)^{r-i}C_i\binom{r}{i}(\frac{(i!)^{\frac{2}{i}}}{i})^{ni}|$. $\alpha(i) := \frac{(i!)^{\frac{2}{i}}}{i}$ is a monotonically increasing function of i for $i > 2$, this can be directly verified using Stirling's approximation for large i, since $\frac{(i!)^{\frac{2}{i}}}{i} \in \frac{O(i^2)}{i} \in O(i)$. For small i, we verified this by numerical simulation. Furthermore, $\alpha(i) \geq 1$ for $i \geq 1$. Therefore $|(-1)^rB + \sum_{i=1,\dots,r}(-1)^{r-i}C_i\binom{r}{i}(\frac{(i!)^{\frac{2}{i}}}{i})^{ni}| \in O((\alpha(r))^{nr}) \in O(\beta(r)^n)$, where $\beta(r) := \alpha(r)^r = \frac{(r!)^2}{r^r}$. Note that $\beta(r) > 1$ for $r > 2$. Therefore

$$|\mathbf{E}(Y^r)| \in \frac{O(n^r(\frac{n}{e})^{nr}\beta(r)^n)}{m^{nr}}.$$

Replacing $n! \approx (\frac{n}{e})^n\sqrt{2\pi n}$ in σ_N^r, then performing straightforward simplifications, we obtain

$$\frac{N|\mathbf{E}(Y^r)|}{\sigma_N^r} \in O\left(\frac{\beta(r)^n}{N^{\frac{r}{2}-1}}\right).$$

Noting, as previously mentioned, that $\frac{N\mathbf{E}(|Y|^r)}{\sigma_N^r} \geq \frac{N|\mathbf{E}(Y^r)|}{\sigma_N^r}$ completes the proof. □