

The trainability of photonic quantum circuits

Alexander Makarovskiy^{*†1}, Adam Taylor^{*‡2}, Zhenghao Li², Michael Hanks²,
Aubrey Clark², M. S. Kim², Ian Walmsley³, and William Clements^{§1}

¹ORCA Computing, London W2 6LA, United Kingdom

²Blackett Laboratory, Imperial College London, London SW7 2AZ, United Kingdom

³Oxford Quantum Institute, Clarendon Laboratory, University of Oxford, Oxford OX3 1PU,
United Kingdom

^{*}These authors contributed equally.

[†]alexander.makarovskiy@gmail.com, [‡]adam.taylor18@imperial.ac.uk, [§]wclements@orcacomputing.com.

Abstract

Variational quantum algorithms are a leading approach to near-term quantum computing, but their scalability can be limited by barren plateaus and the sampling cost of resolving small changes in the loss landscape. Here, we study the trainability of passive linear-optical quantum circuits and introduce a framework based on the ratio of sample variance to circuit variance. This ratio determines the number of circuit samples required to resolve local loss differences and gradients to proportional accuracy. We apply this framework to photon-number observables and identify both trainable and non-trainable regimes. Supported by analytic results and a numerically observed polynomial decay of the circuit variance, we find that fixed-order photon-number polynomials require only polynomially many samples as the system size grows, whereas high-order polynomials and observables based on output probabilities generally require exponentially many samples. Within the trainable regime, we further identify classes of observables in which quantum estimation achieves a polynomial speed-up over multiple classical methods. Within this family, neural network observables provide one practical construction that allow measurement outcomes to be efficiently processed into the desired polynomial. These results establish photonic variational quantum computing as a promising platform for near-term applications.

1 Introduction

Current quantum hardware remains in the Noisy Intermediate-Scale Quantum (NISQ) regime [1], with limited qubit counts, connectivity, and noise levels that place celebrated quantum algorithms such as Shor’s algorithm for factoring large numbers beyond reach [2, 3]. Variational quantum computing is a leading approach for utilising such devices towards applications, encompassing variational quantum algorithms (VQAs) and quantum machine learning (QML) protocols [4, 5]. Variational circuits have found applications in optimisation [6, 7], quantum chemistry [8, 9], machine learning [10, 11], linear algebra [12], state preparation [13, 14] and quantum simulation [15, 16].

The optimism around VQAs and QML has been tempered by a series of no-go results establishing fundamental obstacles to training parameterised quantum circuits. Foremost among these is the barren plateau (BP) phenomenon: a guarantee that, for a sufficiently expressive and randomly initialised circuit, the optimisation landscape is almost entirely flat away [17]. There can still exist a small number of exponentially concentrated minima, referred to as narrow gorges [17, 18], but the probability of finding one is exponentially small. Training over such a landscape requires exponential precision to resolve gradients, followed by a difficult search for the rare regions where pronounced minima exist [18, 19]. These problems can be made worse by the choice of observable [20], excess entanglement [21, 22] or noise in real quantum hardware [23]. A substantial body of work has gone into diagnosing barren plateaus in gate-based regimes [24, 25], and considerable effort has since gone into identifying circuit families and methods that avoid them [22, 26–29]. However, many of the architectures shown to be barren-plateau-free have also turned out to be efficiently classically simulable [30–32]. These results have made it challenging to scale experimental demonstrations of variational algorithms past a few tens of qubits.

In contrast, recent experiments with photonic quantum circuits have successfully demonstrated quantum machine learning tasks at a scale significantly beyond the limitations implied by the theory around barren plateaus. In [33], the authors successfully encode pixel data from the MNIST handwritten-digit dataset into the Hilbert space spanned by a quantum photonic system with 8176 modes and over 2000 photons and perform image classification. In [34], the authors perform quantum reservoir computing with a 400-mode photonic circuit and several hundreds of photons. These results warrant further investigation into how photonic circuits differ in trainability.

However, the previous literature on investigating quantum circuit trainability is not directly applicable to photonics. Near-term photonic architectures often rely on parametrisable linear-optical circuits with quantum state inputs. While not universal for quantum computation, sampling from such a circuit is classically hard despite its relative experimental simplicity [35–38]. This architecture thus provides a promising platform for demonstrating quantum advantage. The natural configurability, absence of decoherence, and high operation speeds of photonic circuits also make them naturally suited for implementing variational protocols. However, the highly constrained gate set, the fact that photonic circuits process particles occupying modes rather than a register of qubit subsystems, and observables which do not naturally map onto qubit-based equivalents mean that practical analyses of the trainability of photonic VQAs require explicit consideration.

Here, we shed light on the trainability of photonic quantum circuits. We define the notions of circuit and sample variance which govern the magnitude of circuit gradients, and sample complexity of estimating them. While natural bounds on observables in gate-based regimes make the identification of circuit variance through barren plateaus often sufficient for diagnosing trainability, we will show that many photonic observables require careful consideration of both variances. We motivate the ratio between them as a metric for trainability, and use this to identify examples of natural photonic observables that are untrainable. We then focus on a different class of observables consisting of linear sums

of products of photon numbers. We show that this class of observables can exhibit polynomially scaling training complexity with increasing system size, but can be efficiently simulated classically. Nonetheless, within this trainable regime we find families of observables for which we can obtain a polynomial speed-up over a different classical methods. We also introduce a neural network observable as an efficient way of implementing a these photon-number polynomials (PNPs). We illustrate these results in both simulation on relevant regimes and on real hardware.

Our results establish that photonic quantum circuits with certain classes of observables can scale to large sizes while maintaining trainability, and also polynomial speed-up over a set of current classical simulation methods. This is promising for finding practical quantum advantage with recently developed applications, ranging from machine learning [33, 39, 40], reservoir computing [34], to generative modelling [41, 42].

2 Background

2.1 Variational photonic circuits

We consider the general variational quantum computing setup of an initial quantum state ρ , a parametrised quantum circuit $\hat{U}(\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$, and measurement of an observable $\hat{O}$. The objective of VQAs or QML models is to find parameters $\boldsymbol{\theta}$ minimising the value of the associated loss function, defined as the expectation value

$$\ell(\boldsymbol{\theta}) = \text{tr}[\hat{O} \hat{U}(\boldsymbol{\theta}) \rho \hat{U}^\dagger(\boldsymbol{\theta})]. \quad (1)$$

More general loss functions can be constructed by considering multiple input states or observables. We will focus on the simple case of Eq. (1), as results in this regime can be extrapolated to the more general cases. Additionally, while parametrised photonic circuits may incorporate mid-circuit measurements and feedforward operations [40, 43–45], we focus exclusively on the dynamics of passive linear optics.

Parameterised quantum circuits are generally trained through a hybrid quantum–classical optimisation loop: a quantum device is used to estimate the loss function at various parameter settings, often to build gradient estimates. These estimates are then supplied to a classical optimiser which updates the parameters towards reducing the loss function. Repeating this procedure iteratively trains the parameters of the circuit towards a loss-minimising regime. Although this framework is largely independent of the underlying hardware platform, the majority of existing work has focussed on qubit-based architectures. In Fig. 1, we illustrate a photonic quantum circuit as part of a variational training loop.

We now turn to the photonic setting that is the focus of this work. We study variational quantum protocols implemented on photonic platforms, with parametrised circuits given by passive linear optical networks (PLONs). We restrict our attention to input states with a fixed total photon number.

An M -mode photonic system is described by the creation and annihilation operators

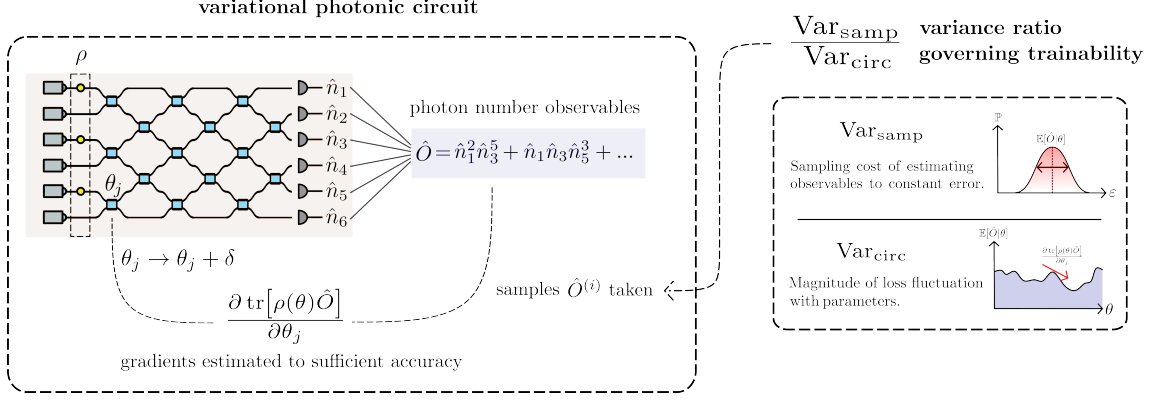


Figure 1: We illustrate the class of VQAs considered in this paper. The ratio between two notions of variance, sample and circuit variance, give a measure of the number of samples needed to estimate accurate gradients to train a photonic circuit with respect to a specific observable. Circuit variance governs the fluctuations of the expected value of an observable with respect to circuit parameters, and exponentially small circuit variance is known as a *barren plateau*. Sample variance determines the number of samples needed to estimate the value of an observable within a given error. We study the trainability of photon-number observables, which exhibit a polynomially scaling variance ratio, making them a suitable for implementing efficiently trainable variational quantum circuits

$\hat{a}_i^\dagger$ and $\hat{a}_i$ respectively for $i \in \{1, \dots, M\}$, satisfying the canonical commutation relations

$$[\hat{a}_i, \hat{a}_j] = [\hat{a}_i^\dagger, \hat{a}_j^\dagger] = 0, \quad [\hat{a}_i, \hat{a}_j^\dagger] = \delta_{ij}. \quad (2)$$

A Fock state has a definite photon number in each mode and is defined by

$$|\mathbf{n}\rangle = |n_1, \dots, n_M\rangle = \prod_{i=1}^M \frac{(\hat{a}_i^\dagger)^{n_i}}{\sqrt{n_i!}} |0\rangle \quad (3)$$

These are eigenstates of the photon-number operator, $\hat{n}_i = \hat{a}_i^\dagger \hat{a}_i$, satisfying $\hat{n}_i |\mathbf{n}\rangle = n_i |\mathbf{n}\rangle$. The n -photon subspace, $\mathcal{H}_{n,M}$ is spanned by all Fock states with total photon number n ,

$$\mathcal{H}_{n,M} = \text{span}\{|\mathbf{n}\rangle \mid \mathbf{n} \in \Phi_M^n\}, \quad \Phi_M^n = \{\mathbf{m} \in \mathbb{N}_0^M \mid \sum_{i=1}^M m_i = n\}, \quad (4)$$

and has dimension $|\Phi_M^n| = \binom{n+M-1}{n}$.

Passive linear optical networks implement transformations that are linear in the creation/annihilation operators and preserve total photon number. Each such transformation is described by a unitary matrix acting linearly on the M modes, i.e. an element of $U(M)$. For a particular $U \in U(M)$, there is an associated optical unitary operator $\hat{U}$ over the entire Hilbert space whose action is

$$\hat{U} \hat{a}_i^\dagger \hat{U}^\dagger = \sum_{j=1}^M U_{ij} \hat{a}_j^\dagger. \quad (5)$$

Since these transformations preserve total photon number, $\hat{U}$ maps each fixed photon

number subspace to itself.

Experimentally, PLONs are constructed from beam splitters and phase shifters. A beam splitter provides a programmable coupling between two modes, controlling how much amplitude is exchanged between them. A phase shifter implements a per-mode tunable phase. Arbitrary linear unitary transformations over M modes can be implemented with a circuit of $M(M-1)/2$ beam splitters with a final layer of phase shifters [46–48].

However, universality over $U(M)$ only provides a universal n -photon *linear-optical* unitary group, which is a proper subgroup of the full unitary group on the n -photon Hilbert space $\mathcal{H}_{n,M}$ [49]. The induced dynamics are restricted to only a subspace of $\mathcal{H}_{n,M}$ decided by the initial state. As such, parametrised PLONs with Haar-random unitary circuit drawn from $U(M)$ are less expressive than circuit-based model of quantum computing, where a random unitary can instead be drawn from the Haar measure over the whole Hilbert space. Despite these restricted dynamics, sampling from the output distribution of a linear-optical circuit, a task known as boson sampling, remains hard to simulate classically [35].

A natural observable to consider for photonic circuits is a product of photon numbers across modes [50], i.e. observables of the form:

$$\hat{M}_{\boldsymbol{\alpha}} = \bigotimes_{i=1}^K \hat{n}_i^{\alpha_i}, \quad (6)$$

We call the subset of $K \leq M$ modes the observable acts on its *support*, and $|\boldsymbol{\alpha}| = \sum_i \alpha_i$ it’s *order*. For some state ρ , the expected value of the observable, $\text{tr}[\hat{M}_{\boldsymbol{\alpha}}\rho]$, is a moment of the photon-number distribution across the modes. We will refer to observables of the form $\hat{M}_{\boldsymbol{\alpha}}$ as photon-number monomials (PNMs).

Any observable diagonal in the photon number basis can be written as a linear sum of PNMs. As such, a study of the trainability of PNM observables generally tells us about the trainability of photonic circuits with respect to arbitrary objective functions.

2.2 Variances as a measure of trainability

Training the parameters of a quantum circuit requires estimating the gradients of the loss function with respect to a given circuit parameter. Generally, the calculation of these gradients analytically is intractable. Practically, we need to sample the quantum circuit at shifted parameter settings, and use estimates of the loss to construct an estimate of the gradient. Throughout this work, we say a circuit–observable pair is *trainable* if its gradients can be resolved to within proportional error with the number of samples required to estimate them growing at most polynomially with the size of the circuit. Conversely, if the number of samples needed to resolve gradients grows exponentially in system size, we say that the choice of circuit is not trainable. In what follows we phrase the trainability requirements in terms of resolving the loss itself; in Appendix B we show that, under mild assumptions, this is sufficient to draw the same conclusions for gradients [18].

This notion of trainability is local: given a parameter setting $\boldsymbol{\theta}$, it asks whether we can efficiently identify an update $\delta\boldsymbol{\theta}$ such that, with high probability, the new parameters

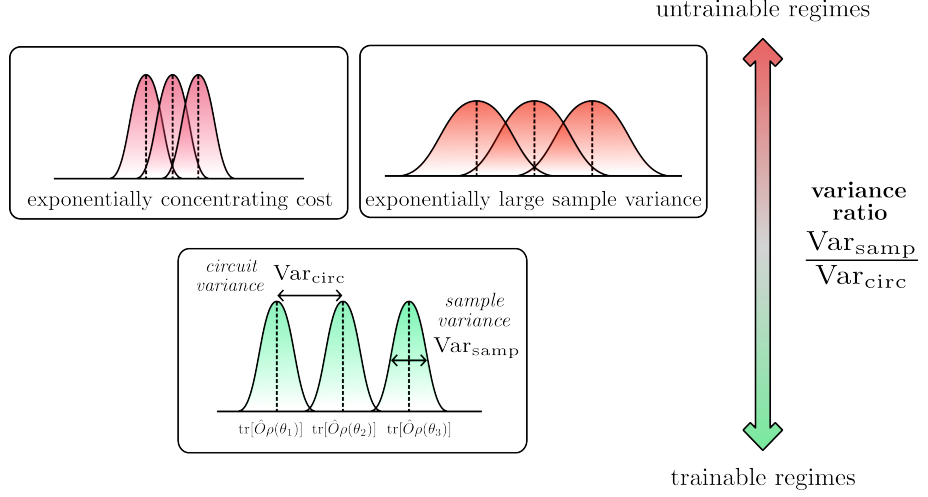


Figure 2: We visualise the distribution of a finite-sample estimate of a fixed observable at different parameter settings. For regimes with a high variance ratio, two observable estimates from different parameter settings will likely not be sufficiently distinct to resolve for which parameter set the ground-truth loss was smaller, due to the overlap in the distributions of the estimates. This can arise either through the means of the observable at different parameter settings being exponentially concentrated (top left), or through the variance of sampling the observable being exponentially large (top right).

satisfy $\ell(\boldsymbol{\theta} + \mathbf{e}_j \delta \theta_j) \leq \ell(\boldsymbol{\theta})$ for all j . In other words, we ask only whether the estimated loss function and gradients contain enough information to make a loss-decreasing parameter update along each parameter direction. We do not address the global optimisation landscape, including the presence of poor local minima, the number of optimisation steps required for convergence, or the training rate needed to avoid overshooting minima.

The first quantity determining the sampling cost is the variance of the measurement outcomes themselves. For a given parameter setting $\boldsymbol{\theta}$, a single run of the circuit yields a measurement outcome of the observable $\hat{O}$ applied to the output state $\rho(\boldsymbol{\theta}) = \hat{U}(\boldsymbol{\theta})\rho\hat{U}^\dagger(\boldsymbol{\theta})$. This distribution has variance $\text{Var}[\hat{O}|\boldsymbol{\theta}] = \text{tr}[\hat{O}^2\rho(\boldsymbol{\theta})] - \text{tr}[\hat{O}\rho(\boldsymbol{\theta})]^2$. The expectation value is estimated by averaging the outcomes of repeated measurements, with the estimation error decreasing as the number of samples increases. The rate of this decrease is governed by the measurement variance $\text{Var}[\hat{O}|\boldsymbol{\theta}]$ such that a broader distribution of outcomes requires more samples to achieve a given precision. Since we are interested in the behaviour of circuits under random initialisation, we average over the distribution of parameters, and study the *sample variance* $\text{Var}_{\text{samp}}(\hat{O}) := \mathbb{E}_{\boldsymbol{\theta}} [\text{Var}[\hat{O}|\boldsymbol{\theta}]]$. This is the expected variance of measurement outcomes when estimating the loss at a typical parameter setting.

The sample variance alone does not tell us how many samples are enough for resolving loss accurately. We need to know the error to which it is sufficient to estimate. The *circuit variance* $\text{Var}_{\text{circ}}(\hat{O}) := \text{Var}_{\boldsymbol{\theta}} [\ell(\boldsymbol{\theta}; \hat{O}, \rho)]$ measures how much the ground-truth loss fluctuates as the circuit parameters are varied, and hence the typical magnitude of the loss differences we must distinguish. In order to resolve a loss-minimising direction, the error of our estimates must be proportional to the magnitude of these differences: the fluctuations of the error must be smaller than the signal of the loss we wish to estimate,

otherwise the estimates carry no information about which parameter settings decrease the loss. The scaling of the circuit variance therefore sets the precision to which the loss must be estimated.

Together, the sample variance sets the cost of achieving a given estimation error, and the circuit variance sets the error we need to achieve. The number of samples needed to estimate observables to resolve a loss minimising direction is then governed by the ratio of these two quantities, $\text{Var}_{\text{samp}}/\text{Var}_{\text{circ}}$. In Appendix A we make this precise, showing that a number of samples scaling with this ratio is sufficient to resolve loss differences at different parameter settings, and hence gradients, to within proportional error. We therefore take this *variance ratio* as our measure of trainability. Circuit–observable pairs whose variance ratio grows at most polynomially with system size are efficiently trainable in the sense that we can efficiently update the parameter in the correct direction. In contrast, an exponentially growing ratio demands exponentially many samples. Fig. 2 illustrates the two ways in which the ratio can become large, namely via the loss values at different parameter settings concentrating exponentially around their mean, or the spread of the estimates around each loss value growing exponentially. In either case, estimates of the loss at different parameter settings overlap too strongly to resolve which is smaller.

For gate-based circuits, natural bounds on the observable magnitude, such as a unit eigenspectrum, bound the sample variance to scale at most polynomially [20]. In this setting the numerator of the variance ratio is benign, and it suffices to study the circuit variance alone. An exponentially decaying circuit variance, known as a *barren plateau* [17, 18], then implies that exponentially many samples are needed to resolve exponentially small gradients, and its absence is sufficient to expect efficient training. These assumptions do not generally hold for photonic circuits. Natural photonic observables, such as products of multiple photon numbers, are unbounded and can take exponentially large values in the number of photons, so the sample variance itself can grow exponentially even in the absence of a barren plateau. In Sec. 3.2 we make this concrete, studying photonic observables that fail to be trainable through either an exponentially large sample variance, or exponentially concentrating circuit variance. Both variances must therefore be considered to conclude trainability in photonic circuits.

3 Trainability of photonic observables

Motivated by the trainability conditions established in Sec. 2.2, we derive expressions for the circuit variance and sample variance of photon-number observables in the Haar-average regime. To do so, we compute the Haar-averaged first and second moments of photon-number observables, from which both variances follow directly. In Sec. 3.1, we present the analytic results for the first moment and second moment. In Sec. 3.2, we apply these equations to calculate the trainability ratio of different photonic observables. We start with photon-number monomials and polynomials, showing that the complexity of training scales with the order of the polynomial, and identifying fixed-order polynomials as trainable. We then discuss neural network observables as a natural implementation of photon-number polynomials. In Sec. 3.2 we present negative results about the non-

trainability of probability output estimating observables. We present the tools developed to produce these results and derivations in the Appendix.

3.1 Variance calculations for photon-number observables

In this section, we present the analytic expressions for the first and second moments used in the variance ratio, for both PNMs and PNPs. Writing $\mu_k(\hat{O}) = \mathbb{E}_U [\text{tr}[\hat{O} U \rho U^\dagger]^k]$ for the k^{th} Haar-average moment of the loss, the circuit and sample variances can be written as differences of its first and second moments,

$$\text{Var}_{\text{samp}}(\hat{O}) = \mu_1(\hat{O}^2) - \mu_2(\hat{O}), \quad (7)$$

$$\text{Var}_{\text{circ}}(\hat{O}) = \mu_2(\hat{O}) - \mu_1(\hat{O})^2. \quad (8)$$

The trainability of an observable–initial state pair therefore reduces to computing the first two moments over the Haar measure.

We consider μ_1 first. The Haar-averaged output state of a linear optical network acting on an n -photon input state is the maximally mixed state over the n -photon subspace – see Appendix C for details. As a result, the Haar-averaged first moment of an observable reduces to its value over an equally weighted sum over all possible output patterns. In Appendix D, we prove the following about photon number monomials;

Theorem 1 (First moment of photon-number monomials). *Let $\hat{M}_\alpha$ be a PNM with $\alpha = (\alpha_1, \dots, \alpha_K)$, and let ρ be an input state with total photon number n . The expectation value of the observable in the Haar-average regime is given by,*

$$\mu_1(\alpha; n) = \sum_{P=K}^{|\alpha|} \frac{\binom{n+M-1}{n-P}}{\binom{n+M-1}{n}} \sum_{\mathbf{p} \in \Phi_K^P} \left(\prod_{i=1}^K p_i! S(\alpha_i, p_i) \right), \quad (9)$$

where $S(\alpha_i, p_i)$ are Stirling numbers of the second kind. Furthermore, if the photon density is constant with $n = \nu M$, then in the large- M limit this approaches a product of ordered Bell polynomials, $F_a(x) = \sum_{p=0}^a p! S(a, p) x^p$, up to $O(1/M)$ corrections,

$$\mu_1(\alpha; n = \nu M) \xrightarrow{M \rightarrow \infty} \left(\prod_{j=1}^K F_{\alpha_j}(\nu) \right) + O(1/M). \quad (10)$$

Crucially, if K and $|\alpha|$ do not scale with M , then the first moment asymptotically approaches a constant value.

Any PNP first moment can be written as a sum over PNM first moments via $\mu_1(\sum_\alpha c_\alpha \hat{M}_\alpha) = \sum_\alpha c_\alpha \mu_1(\alpha)$ and hence Theorem 1 is applicable to polynomials too. We can now analyse the scaling of the first moment in terms of the order of the monomial and the circuit size. For a constant-density regime $n = \nu M$, where the first moment is given by a product of ordered Bell polynomials in Eq. (10).

In Appendix D.2, we show that for large $|\alpha|$ the first moment asymptotically scales as

$$\mu_1(\alpha; n = \nu M) \rightarrow \prod_{i=1}^K F_{\alpha_i}(\nu) \sim \prod_{i=1}^K \frac{\alpha_i!}{(1 + \nu) \ln(1 + 1/\nu)^{\alpha_i+1}}, \quad (11)$$

This grows factorially in the order and exponentially in the number of modes the PNM is supported on. In Fig. 10, we plot $\mu_1(\alpha)$ against $|\alpha|$ for fixed K , and find that it does indeed grow superexponentially.

Calculation of the second moment requires the second-order Haar moments of the output state. We apply the techniques developed by Mhiri *et al* [51], which decompose the second moment over the irreducible representations of $U(M)$ within the n -photon subspace to give [51]

$$\mu_2(\hat{O}; \rho) = \sum_{k=0}^n \frac{1}{d_k^{(n)}} \|\rho_k^{(n)}\|_2^2 \|\hat{O}_k^{(n)}\|_2 \quad (12)$$

Here, $\hat{X}_k^{(n)}$ is a projector onto the k^{th} irreducible representation of $U(M)$ and $\|\bullet\|_2^2 = \text{tr}[\bullet^\dagger \bullet]$ is the Hilbert-Schmidt norm. In Appendix E, we present or derive the projected norms for Fock states, entangled states, PNMs and PNPs. These values can be efficiently computed using discrete convolutions, allowing us to numerically study the scaling of the second moments.

Unlike the first moment results, analytic asymptotic formulae for the second moments are less clear. Through the law of total expectation and Jensen's inequality, we can guarantee that $\mu_1(\alpha)^2 \leq \mu_2(\alpha) \leq \mu_1(2\alpha)$; by Eq. (10) we know that for fixed $n = \nu M$ and $K, |\alpha| \ll M$ then the second moment is necessarily between two system-size independent constants, and thus itself approaches a constant. We numerically observe that $\mu_2(\alpha)$ approaches the lower end, with $\mu_2(\alpha) - \mu_1(\alpha)^2$ appearing to decay slowly with system size. We conjecture that this decay is polynomial with respect to M . In Fig. 4, we plot $\text{Var}_{\text{circ}}(\hat{n}^\alpha)$ against M for different choices of α with initial state $|\mathbf{1}, \mathbf{0}\rangle$ and constant density ν . We find all PNM first moments peak at some M_{max} before slowly decaying, and the value of M_{max} grows with α . From the numerics alone we cannot perfectly distinguish the exact nature of the slow decay component, but we conjecture it is of the form

$$\text{Var}_{\text{circ}}(\alpha) \sim \frac{\mu_1(\alpha)^2}{\text{poly}(M)}. \quad (13)$$

This is well justified by Fig. 3 where we see non-linear decay in a log-plot of the circuit variance across a series of low order PNMs. In Fig. 4, we again see a very slow decay of the circuit variance across a variety of photon-number monomials $\hat{n}^\alpha$. The analytic formulae above, together with these numerical observations, characterise the quantities of sample and circuit variance governing trainability for photon number monomials and give insight into more general photon number polynomials.

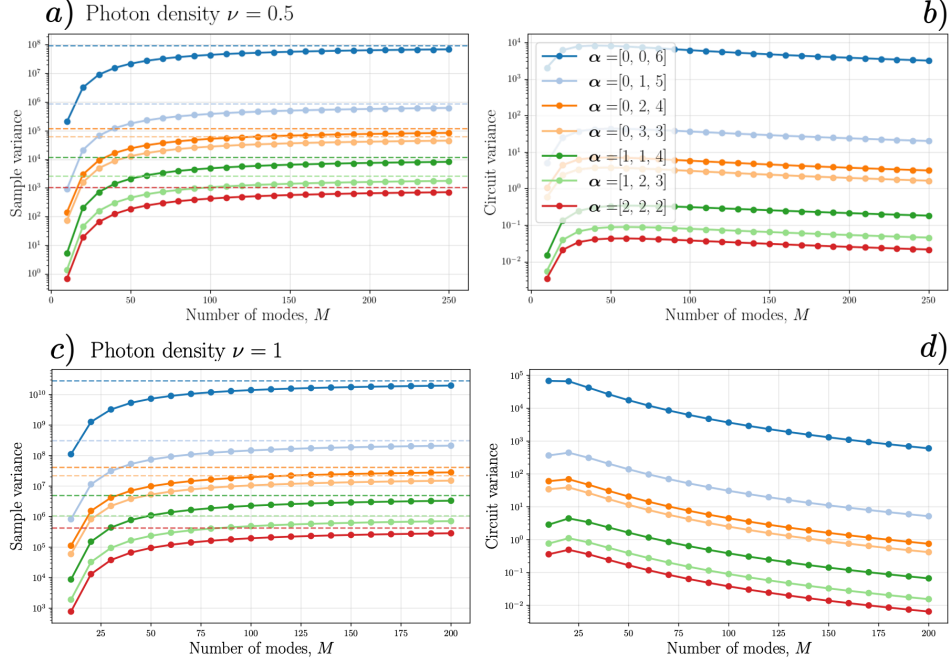


Figure 3: Sample and circuit variances as functions of the number of modes M for PNMs satisfying $|\alpha| = 6$ and $K \leq 3$. Panels (a) and (c) show the sample variance for photon densities $\nu = 0.5$ and $\nu = 1$ respectively. The dotted lines denote the asymptotic limits $\mu_1(2\alpha)|_{M \rightarrow \infty}$, towards which the sample variances converge. Panels (b) and (d) show the corresponding circuit variances. In both cases, the circuit variance decays only very slowly with M . Its visibly non-linear behaviour on the logarithmic vertical scale suggests that the decay is not exponential. We therefore conjecture that the circuit variance instead decays polynomially with M , as stated in Eq. (13).

3.2 Trainable and non-trainable regimes

We apply the results calculated above to analyse the trainability of different photon number polynomial regimes. We also analyse the trainability ratio of output probability estimation observables that have previously been used in the literature. Based on previous tradeoffs that have been observed between trainability and simulability, we comment on the classical simulability of these observables.

3.2.1 Photon-number monomials

The scaling results of Sec. 3.1 allow us to directly calculate the variance ratio of photon-number monomials. We will again consider the regime where $n = \nu M$. Using the observations that the circuit variance decays polynomially, $\text{Var}_{\text{circ}}(\alpha) \approx \mu_1(\alpha)^2 / \text{poly}(M)$, and that the sample variance is dominated by the highest moment, $\text{Var}_{\text{samp}}(\alpha) \approx \mu_1(2\alpha)$, the order scaling of Eq. (11) gives

$$\frac{\text{Var}_{\text{samp}}(\alpha)}{\text{Var}_{\text{circ}}(\alpha)} \approx \text{poly}(M) \cdot \frac{\mu_1(2\alpha)}{\mu_1(\alpha)^2} \xrightarrow{M \rightarrow \infty} \text{poly}(M) \cdot \left(\prod_{i=1}^K \binom{2\alpha_i}{\alpha_i} (1 + \nu) \ln(1 + 1/\nu) \right). \quad (14)$$

Fixed-PNM circuit variance

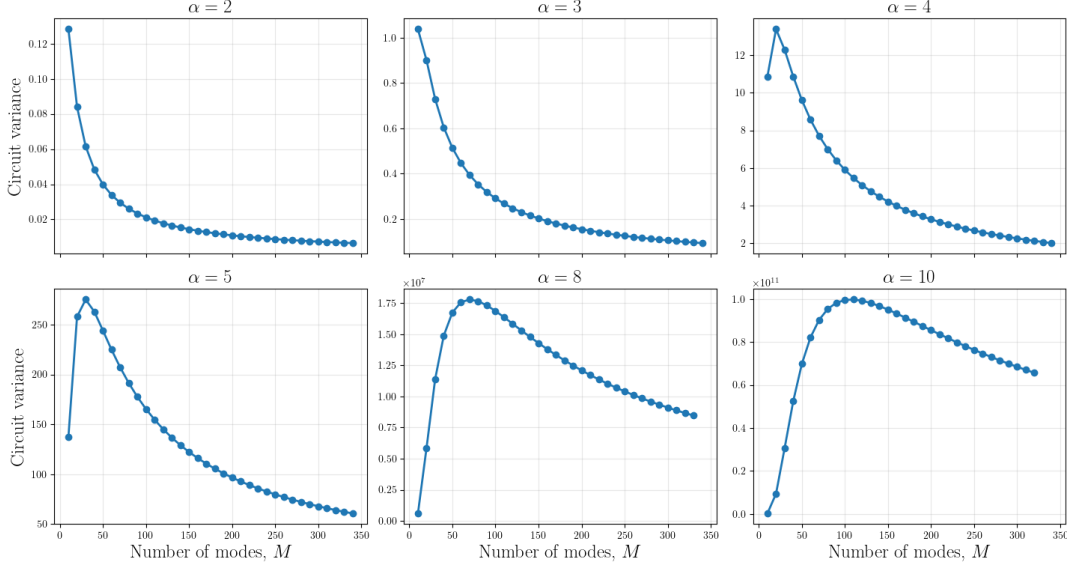


Figure 4: Circuit variance as a function of the number of modes M for observables $\hat{n}^\alpha$ at several values of α . The input state is $|\mathbf{1}, \mathbf{0}\rangle$ with photon density $\nu = 0.5$. For each α , the circuit variance initially increases before entering a regime of very slow decay, with the maximum shifting to larger M as α increases. Although deriving the asymptotic decay analytically is challenging, the numerical results suggest a polynomial dependence on M , consistent with the scaling analysis in Fig. 3.

The variance ratio scales polynomially in system size, with a factor that grows exponentially with the total order and support of the photon number monomial observable.

For a fixed-order monomial at constant photon density, the exponential factor in Eq. (14) is a system-size-independent constant, and the variance ratio is strictly polynomial in M . This means fixed-order PNMs exhibit no barren plateaus as the system size scales, and their losses can be resolved to the scale of their circuit fluctuations with polynomially many shots. Therefore, we conclude that variational photonic circuits can be efficiently trained on fixed photon-number monomials.

Monomials whose total order $|\alpha|$ grows linearly with the number of modes M behave differently. The exponential factor in Eq. (14) then grows exponentially with *system size*, so we do not expect such observables to be trainable. For example, for the monomial $\alpha = (1, \dots, 1)$, which takes a product of photon numbers over $K \leq M$, consider how the variances scale. Just looking at the circuit variance, we find a scaling

$$\text{Var}_{\text{circ}}(\alpha) \approx \nu^{2K} / \text{poly}(M). \quad (15)$$

One might think this observable exhibits the opposite of a barren plateau – in the over-saturated regime with $\nu > 1$, the circuit variance grows exponentially fast with M when $K \propto M$. However, the sample variance grows exponentially faster

$$\text{Var}_{\text{samp}}(\alpha) \geq (2^K - 1)\nu^{2K} \quad (16)$$

and hence the variance ratio scales as $\Omega(2^K)$, which is also exponential in M when $K \propto M$. Resolving accurate loss function gradients governed by such observables therefore requires exponentially many samples, and it is not sufficient to only study the circuit variance to identify this. Intuitively, the observable exhibits large fluctuations with parameters, but also has rare large magnitude events contributing to the majority of this, meaning estimating these values through sampling is exponentially hard.

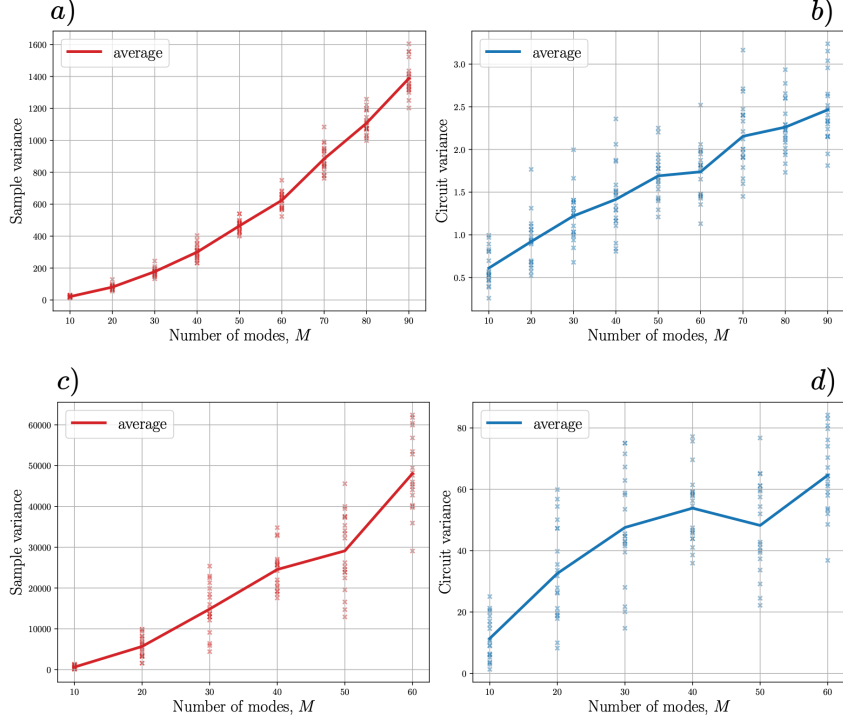


Figure 5: Numerical scaling of the circuit and sample variances for two classes of photon-number polynomials (PNPs). For each value of M , we generate 20 PNPs with independently sampled random coefficients. Individual instances are shown as crosses, while solid lines indicate the corresponding averages. In panels (a) and (b), the PNPs contain all photon-number monomials (PNMs) satisfying $|\alpha| \leq 2$, with coefficients sampled uniformly from $[-1, 1]$. Since the number of terms scales as $|A| \sim O(M^2)$, the average sample variance in (a) exhibits approximately quadratic growth. The circuit variance in (b) grows approximately linearly, indicating that the increasing number of terms compensates for the decay of the circuit variance of the constituent PNMs. In panels (c) and (d), we consider PNPs of order $|\alpha| \leq 4$, each containing $|A| = 5M$ randomly selected PNMs with coefficients sampled uniformly from $[-1, 1]$. The average sample variance grows approximately linearly, consistent with the predicted scaling $O(|A|)$. The circuit variance also increases with M , although it shows a weak tendency towards a plateau at larger system sizes. In both cases, the absence of a decaying circuit variance and polynomially scaling sample variance indicates that the corresponding loss functions remain trainable over the system sizes considered.

3.2.2 Photon-number polynomials

A single photon-number monomial is a restricted class of observable, and their product form makes them susceptible to efficient classical simulation methods [52]. We therefore fo-

cus on photon number polynomials, expressed as $\hat{C} = \sum_{\alpha \in A} c_{\alpha} \hat{M}_{\alpha}$. Indeed, any diagonal observable in the photon number basis can be written as a finite PNP.

The PNP first moment $\mu_1(\hat{C})$ is just given by the sum over PNM first moments weighted by c_{α} . The first moment of the squared PNP is given by $\sum_{\alpha, \beta \in A} c_{\alpha} c_{\beta} \mu_1(\alpha + \beta)$ with each term again given by Theorem 1. Meanwhile, to compute the second moment requires computing $\|\hat{C}_k^{(n)}\|_2^2$, which we carry out in Appendix E.3. We will first discuss the sample variance.

Without making any assumptions on the coefficients, the sample variance is bounded by the constituent monomials using the Cauchy-Schwarz inequality, $\mu_1(\alpha + \beta) \leq \sqrt{\mu_1(2\alpha) \mu_1(2\beta)}$. This gives the bound

$$\text{Var}_{\text{samp}}[\hat{C}] \leq \mu_1(\hat{C}^2) \leq \left(\sum_{\alpha \in A} |c_{\alpha}| \sqrt{\mu_1(2\alpha)} \right)^2 \leq |A|^2 \max_{\alpha \in A} c_{\alpha}^2 \mu_1(2\alpha) \quad (17)$$

where $|A|$ is the cardinality of the set A , equal to the number of PNM terms present in the PNP. If the polynomial has a maximum constant order, and the coefficients are bounded, each $\mu_1(2\alpha)$ is a system-size independent constant, so a PNP with polynomially many terms can be estimated to fixed additive precision $\varepsilon_{\text{additive}}$ with a number of shots scaling as $O(|A|^2)$ – more detail is given in Appendix D.3.

For typical coefficients we can strengthen this bound. Consider a set of random coefficients drawn from a distribution satisfying $\mathbb{E}[c_{\alpha}] = 0$ and $|c_{\alpha}|$ being independent of M for all α . In Appendix D.3, we show that for such coefficients,

$$\mathbb{E}_{\{c_{\alpha}\}} \left[\text{Var}_{\text{samp}}(\hat{C}) \right] \leq \sum_{\alpha \in A} \mathbb{E}_{\{c_{\alpha}\}} [c_{\alpha}^2] \mu_1(\alpha) \sim O(|A|) \quad (18)$$

In this case, the expected sample variance improves from $O(|A|^2)$ to $O(|A|)$. Similarly, the same off-diagonal cancellation takes place at the level of the circuit variance, leading to

$$\mathbb{E}_{\{c_{\alpha}\}} \left[\text{Var}_{\text{circ}}(\hat{C}) \right] = \sum_{\alpha \in A} \mathbb{E}_{\{c_{\alpha}\}} [c_{\alpha}^2] \text{Var}_{\text{circ}}(\alpha). \quad (19)$$

As such, the ratio of the expected variances, $\mathbb{E}_{\{c_{\alpha}\}}[\text{Var}_{\text{samp}}(\hat{C})] / \mathbb{E}_{\{c_{\alpha}\}}[\text{Var}_{\text{circ}}(\hat{C})]$, becomes a weighted average of the constituent monomial averages. A typical fixed-order PNP therefore inherits the $\text{poly}(M)$ variance ratio of Eq. (14) and continues to be trainable. In contrast, PNPs with a non-negligible set of monomials that scale with system size inherit their exponentially scaling variance ratio and are not trainable. Beyond these stochastically chosen PNPs, we note that highly structured PNPs can have coefficients that lead to cancellations that suppress either variance.

In Fig. 5, we plot the sample variance and circuit variance for classes of low-weight PNPs; in (a) and (b), we consider all summing over all PNMs with weight $|\alpha| \leq 2$ and coefficient $c_{\alpha} \sim \text{Uniform}[-1, 1]$, while in (c) and (d), we take a set of $|A| = 5M$ PNMs with weight ≤ 4 . We see the sample variance for both PNMs tracks has the predicted scaling of $O(|A|)$, going as $|A| \sim O(M^2)$ in (a) and $|A| \sim O(M)$ in (c).

The circuit variances also increase over the range of system sizes considered, despite

the circuit variance of each constituent PNM decreasing with M , as shown in Figs. 3 and 4. This behaviour can be understood as the increasing number of terms compensating for the polynomial decay of the individual PNM circuit variances. For example, in panel (b), each constituent PNM has a circuit variance that decreases with M , with similar behaviour to the $\alpha = 2$ curve in Fig. 4. However, Eq. (19) shows that the coefficient-averaged circuit variance is obtained by summing the contributions from all $|A|$ constituent PNMs. Over the plotted range, the growth in $|A|$ therefore outweighs the decay of the individual contributions, resulting in an overall increase in the circuit variance.

Comparison to classical estimation. Previous literature has strongly suggested that efficiently trainable variational quantum circuits must admit efficient classical simulation, trading quantum advantage for trainability [30]. We therefore study the complexity of classically spoofing the training of the trainable PNPs above, using one of the best-known general-purpose classical algorithms for observable estimation in linear optics, the Lim–Oh algorithm [52]. The algorithm efficiently estimates the expectation values of observables in product form over product input states, requiring a $O(\|\hat{O}\|_2^2/\varepsilon_{\text{additive}}^2)$ time-complexity to produce an estimate to additive error $\varepsilon_{\text{additive}}$. This is the classical analogue to the sample complexity of a quantum circuit. A generic PNP admits no such product form, so the algorithm must instead be applied term by term, estimating each constituent monomial and summing them, or via Monte-Carlo importance sampling over the most heavily weighted terms. In Appendix D.4, we show that both strategies carry a complexity of $O(|A|^2/\varepsilon_{\text{additive}}^2)$. The Lim-Oh simulation method cannot exploit the cancellation of coefficients in the sample variance ratio as it must be applied to the constituent product observables individually. A quantum device sampling the polynomial observable directly does: for balanced coefficients the sample variance scales with $|A|$ by equation (18). This suggests a quadratic speed-up using quantum sampling versus Lim-Oh simulation methods for the task of estimating the loss function.

Since $|A|$ can grow rapidly with the maximum order of the polynomial, this separation can be relatively large even at modest system size. Moreover, in Sec. 3.2.3 we introduce neural network observables as an efficient means of implementing PNPs with exponentially many terms without ever expanding their constituent coefficients, a regime in which the term-wise classical simulation methods can be disadvantaged further.

Other classical simulation methods generate samples that reproduce the statistics of an ideal boson-sampling distribution up to the a^{th} order correlators [53, 54]. Therefore, for PNP observables satisfying $|\alpha| \leq a$, these samples reproduce the same expectation values relevant to $\hat{C} = \sum_{\alpha \in A} c_{\alpha} \hat{M}_{\alpha}$. The number of generated samples required to estimate this expectation value therefore has the same scaling as the quantum sample complexity in Eq. (18) or Eq. (19), so there is no quantum sampling advantage with this approach. However, generating each classical sample has a classical cost $O(M^a)$ for $a = \max_{\alpha} \{|\alpha|\}$ [54], leaving space for a potential polynomial quantum speed-up for sufficiently chosen a . Moreover, in their current form these methods require modification to accommodate photon-number-resolving observables and are formulated for Gaussian input states. Adapting them to the

photonic VQA setting considered here will therefore be important for determining whether genuine polynomial quantum runtime advantage can persist within the trainable regime.

We therefore expect photonic variational circuits to exhibit a polynomial speed-up in the runtime of variational protocols on a large class of PNP observables when compared with the currently available classical algorithms. Conversely, the fact that the classical estimate remains polynomial-time motivates a train-on-classical, deploy-on-quantum regime [41, 55]. Here, the pre-training of a circuit can be performed classically with some overhead (either $|A|$ with the Lim-Oh approach [52] or $M^{|\alpha|}$ with the boson sampling emulator approach [53, 54]).

3.2.3 Neural network observables

A natural way of implementing photon-number polynomials is to consider a neural network as an observable of a variational quantum circuit. Each photon number sample is passed in as the input to a multi-layered neural network. We take the activation functions in the neural network to be quadratic. This way, we can simply write out the action of the neural network as a multivariate polynomial, allowing us to apply previous results.

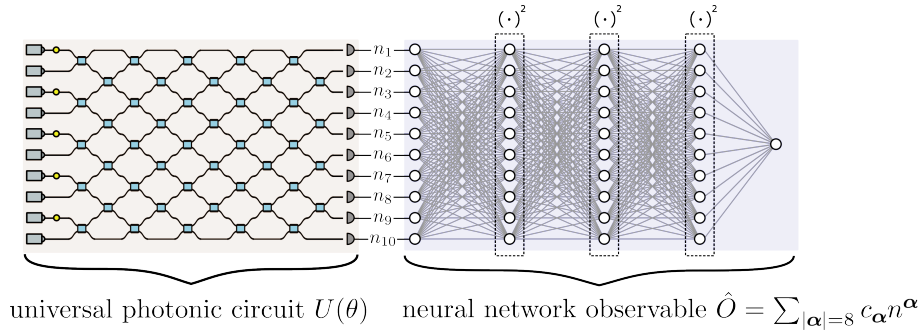


Figure 6: We show a 3-layer neural network observable implementing a photon-number polynomial operation. In red, we highlight a path that adds weight to one of the monomial terms.

Consider a network with L linear layers, each interleaved with a quadratic activation function, whose final layer has a single output neuron. Its action can be written as

$$\text{NN}(\mathbf{n}) = \sum_{|\alpha| \leq 2^L} c_{\alpha} \mathbf{n}^{\alpha}. \quad (20)$$

In words, the NN observable is equal to a PNP including all possible PNM terms up to order 2^L . There are $\binom{2^L+M}{M}$ terms in this polynomial. For $2^L \ll M$, this scales as $\approx M^{2^L}/(2^L)!$. The coefficients c_{α} are determined by the weights and biases of the linear layers.

While the polynomial representation grows exponentially in size, applying the forwards pass of a neural network to a sample has complexity $O(LM^2)$: we need to multiply the sample by L size $M \times M$ matrices, and apply quadratic activation functions to the intermediary terms. This is a key advantage of the NN representation: it sidesteps the need to enumerate the full polynomial explicitly. By contrast, classical observable-estimation

methods that require the observable to be expressed as a sum of product observables must first expand the network into this polynomial form, and then either estimate each term individually or use a weighted estimation scheme. In Appendix D.4, we discuss various classical strategies for estimating polynomial observables.

3.2.4 Output probability estimation

Several proposals for boson sampling applications rely on estimating either a single output probability or a scaled aggregate of output probabilities. Although such observables are of theoretical interest, we show that they cannot be estimated to proportional error by a physical boson sampler with polynomial resources. This rules out their use in variational protocols, and further suggests that previous proposals relying on these observable estimates are unscalable without careful consideration of the magnitudes of probabilities being estimated

Single-probability estimation A natural loss function to consider is one governed by a projection onto a state $|\psi\rangle\langle\psi|$. Indeed, estimating the expectation value $\text{tr}(\rho(\theta)|\psi\rangle\langle\psi|)$ has been proposed as an application of boson sampling towards problems in graph theory [56, 57]. We will show that the trainability ratio of projector observables is exponentially large, illustrating their non-suitability for variational protocols.

As before, we will consider an input state with fixed photon number, and will set $|\psi\rangle = |\mathbf{s}\rangle = |s_1, \dots, s_M\rangle$. In Appendix G.1, we compute the Haar-average moments of the projector observable, from which the variances follow,

$$\text{Var}_{\text{circ}}(|\mathbf{s}\rangle\langle\mathbf{s}|) = (P_2(M, n) - 1)|\mathcal{H}|^{-2}, \quad (21)$$

$$\text{Var}_{\text{samp}}(|\mathbf{s}\rangle\langle\mathbf{s}|) = |\mathcal{H}|^{-1} (1 - P_2(M, n))|\mathcal{H}|^{-1}, \quad (22)$$

where $P_2(M, n)$ is the normalised average collision probability, shown to scale as $\Theta(n + n/M + 1)$ [51]. We get a variance ratio:

$$\frac{\text{Var}_{\text{samp}}(|\mathbf{s}\rangle\langle\mathbf{s}|)}{\text{Var}_{\text{circ}}(|\mathbf{s}\rangle\langle\mathbf{s}|)} = \frac{|\mathcal{H}| - P_2(M, n)}{P_2(M, n) - 1} = \Theta\left(\frac{|\mathcal{H}|}{n + n/M + 1}\right). \quad (23)$$

With $|\mathcal{H}| = \binom{n+M-1}{n}$ growing exponentially in n, M , the difficulty of resolving the difference between two loss functions governed by the projector grows exponentially with problem size, ruling out its use in trainable variational protocols at scale. It may be natural to consider scaling the projector by a constant c to normalise the circuit variance to be constant (indeed, this is done to relate estimated probabilities to permanents [57]). However, the variance ratio suggests that estimating such quantities to proportional precision remains exponentially hard – constant factors that increase the circuit variance also increase the sample variance by the same factor, resulting in no change in the variance ratio. We expect this makes the scaling of applications relying on resolving output probabilities challenging without careful consideration of the circuit and projector.

We note that this is the photonic analogue of a known gate-based circuit result: loss functions measuring fidelity to a target state produce barren plateaus even in shallow circuits [20].

Coarse-grained sampling To avoid the exponentially hard task of resolving a single output probability of a quantum circuit, it has been proposed to estimate the probability of a collection of output measurements, a regime known as ‘coarse-grained sampling’ [58, 59]. We show that, while the sample variance of such observables is indeed constant, the circuit variance is exponentially small. This limits the scalability of such regimes for variational protocols.

Furthermore, proposals for the use of coarse-grained boson sampling for cryptographic applications and proof-of-work schemes relying on resolving which set of output probabilities is larger are likely unscalable without careful circuit and observable selection.

A coarse-grained boson sampling observable is defined as a uniformly weighted sum of the single-outcome projectors as considered above, chosen from a set $S \subseteq \mathcal{H}_{n,M}$,

$$\hat{O}_S = \sum_{\mathbf{s} \in S} |\mathbf{s}\rangle\langle\mathbf{s}|, \quad S \subseteq \mathcal{H}_{n,M}, \quad (24)$$

The expectation value of this observable is the total probability mass assigned to the bin S . There have been several proposals which consider estimating such expectation values to decide which of several bins is most likely as a cryptographic one-way function that is hard to invert, but computable with a boson sampler [60, 61]. Following these proposals, we will consider bins comprising a constant fraction of outcomes, $|S| = |\mathcal{H}|/K$ for constant K .

Sampling the observable $\hat{O}_S$ with a Haar-randomly generated circuit is a Bernoulli variable with success probability $\mathbb{E}_\mu[\mathbb{P}(S|\theta)] = |S|/|\mathcal{H}| = 1/K$. As such, the total Haar-average variance is $\frac{1}{K}(1 - \frac{1}{K})$, and the law of total variance gives the identity,

$$\frac{1}{K} \left(1 - \frac{1}{K}\right) = \text{Var}_{\text{samp}}(\hat{O}_S) + \text{Var}_{\text{circ}}(\hat{O}_S). \quad (25)$$

Sample and circuit variance of a constant-fraction coarse-grained output observable sum to a constant. We show below that the circuit variance is typically exponentially small, so we can conclude the sample variance is $\Theta(1)$.

Unlike for single outcome probabilities, the circuit variance of the coarse-grained observable depends on the choice of bin S , so we compute its average over sets S drawn uniformly at random among the constant-fraction subsets of $\mathcal{H}$. In Appendix G.2, we derive the average circuit variance:

$$\mathbb{E}_S \left[\text{Var}_{\text{circ}}(\hat{O}_S) \right] \approx \frac{1}{K} \left(1 - \frac{1}{K}\right) \frac{P_2(M, n) - 1}{|\mathcal{H}|}. \quad (26)$$

As we mention previously, $P_2(M, n) = \Theta(n + n/M + 1)$ [51]. As $|\mathcal{H}|$ is exponentially growing in n, M , the above result means that only an exponentially small fraction of

choices of sets S can have non-exponentially small circuit variance. As such, almost all choices of output aggregation concentrate around the mean and exhibit a barren plateau.

While such coarse-grained observables have a constant sample variance unlike single output probability estimation, the exponentially small circuit variance means their trainability ratio remains exponentially large: an exponentially large number of samples is needed to resolve the value of such observables with sufficient precision. This does not preclude carefully structured output binning from admitting polynomial variance ratios, but our results show that such binnings constitute an exponentially small fraction of all choices, and identifying them is a nontrivial task that proposals relying on coarse-grained estimates must address.

4 Numerical simulations and photonic experiments

4.1 Single-photon input states

We consider the experimentally relevant regime of training circuits with an input state of $n = M/2$ photons being propagated over M modes. As discussed in Sec. 3.2.1, for such a regime, the number of samples needed to resolve gradients for each parameter scales polynomially in system size for a fixed-order photon-number polynomial.

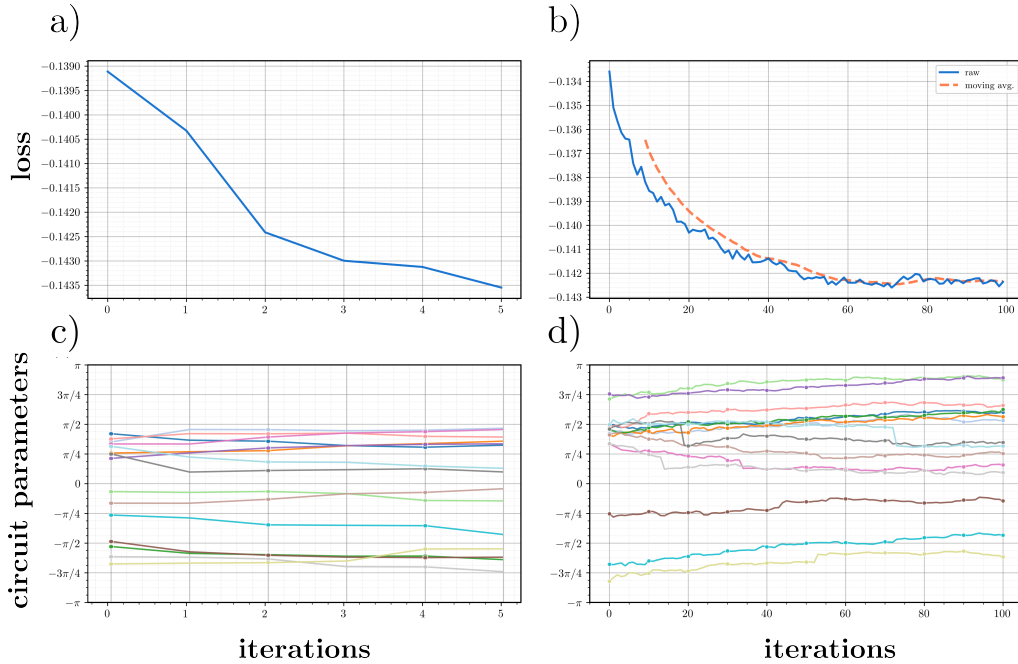


Figure 7: Training of simulated photonic circuits on a 3-layer neural network observable. (a,c) a 13-photon, 26-mode configuration trained for 6 iterations, and (b,d) a 10-photon, 20-mode configuration trained for 100 iterations. 10,000 samples are taken for each gradient estimate using the generalized parameter shift rule (GPSR) taken up to the 8th frequency. (c) and (d) illustrate the evolution of 15 randomly chosen circuit parameters, showing smooth trajectories.

We train simulated photonic circuits arranged in a Clements scheme beam splitter configuration, initialised at a Haar-random unitary. We take the observable to be a ran-

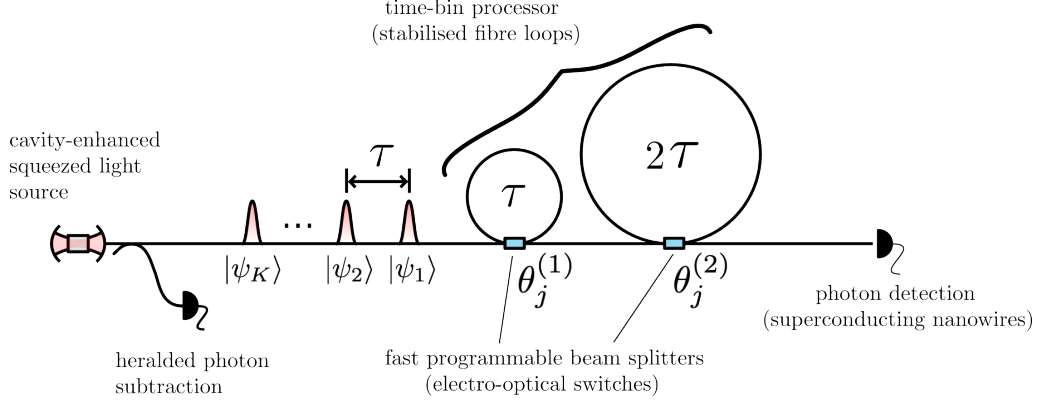


Figure 8: Illustration of the PT-2 photonic system. Squeezed states of light are generated at time intervals separated by delay τ . Samples are accepted when a photon subtraction has been heralded on at least one input squeezed state, resulting in a non-Gaussian input. The input state is interfered between modes through two sequential delay loops. An array of nanowire detectors read out pseudo-photon number resolving measurements.

domly initialised 3-layer neural network, giving a total observable that can be written as a photon-number polynomial with all terms up to order 8. Our results show that a modest sample budget of 10,000 samples per gradient update remains sufficient for training intermediate scale regimes. Due to the prohibitive simulation complexity of photonic circuits, we present training runs for only 20 and 26 modes.

4.2 Experiments on photonic hardware

We demonstrate training photonic hardware with the ORCA PT-2. We illustrate that near-term photonic hardware successfully trains with high accuracy on loss functions given by low-order photon-number polynomials. The ORCA PT-2 is a time-bin encoded photonic quantum processor. It samples from circuits with two cascading layers of beam splitters, corresponding to two optical delay lines. While such circuits are not universal and cannot realise arbitrary Haar-random unitaries, sampling from their output distribution remains classically hard [36, 62]. Moreover, as they are significantly shallower than universal linear-optical circuits, they are generally expected to be easier to train, with training complexity no greater than that of the deeper architectures considered in our analysis. We illustrate a schematic of the hardware in Fig. 8.

We train the PT-2 on a 100-term photon-number polynomial up to order 2, with randomly chosen weights between

$$[-5, 5]$$

. We post-select samples to have a fixed number of photons. Under assumptions of uniform loss, even with Gaussian input states, such a sampling regime can be equivalently described through an input superposition state of fixed photon number. The theory developed in Sec. 3 applies directly to such a regime. In Fig. 9, we see that with a modest budget of 1000 samples per gradient estimate, the PT-2 can efficiently train on a large low-order photon-number polynomial.

This result motivates training near-term photonic system on loss functions governed

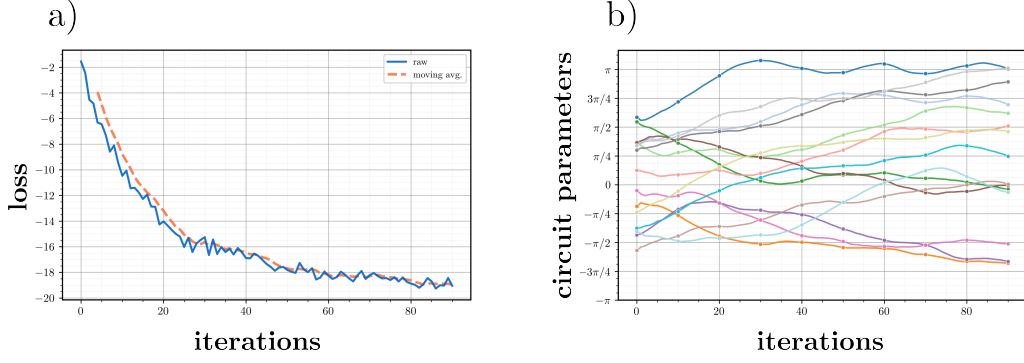


Figure 9: Training of a PT-2 system to minimise loss. The PT-2 is run for 48 modes per sample, with an alternating 1,0 input state of 24 squeezed states generated. Samples are post-selected on having exactly 6 photons present. The loss function is set with respect to minimise a 100-term photon-number polynomial up to order 2, with random weights chosen between $[-5,5]$ for each term. 1000 samples are taken for each gradient estimate, which is done through the parameter shift rule. Figure a) shows loss convergence smoothly towards a local minimum. Figure b) illustrates the smooth training of a randomly chosen set of 15 circuit parameters.

by the minimisation of a low-order photon-number polynomial observable. Indeed, many relevant classes of observables, such as those arising from quadratic unconstrained binary optimisation (QUBO) problems, can be described as order two photon-number polynomials, and previous proposals consider training photonic circuits on such objectives [63, 64]. It remains an open question to quantify the effect of experimental imperfections such as errors and parameter jitter on the trainability of photonic circuits; photon losses are expected to be problematic as systems increase in size, so understanding how post-selecting on error-free outcomes or implementing photonic error mitigation [65, 66] impact total training time will be important. Nonetheless, these preliminary experimental results indicate that NISQ photonic systems can be trained despite imperfections.

5 Discussion and conclusion

In this work, we introduced the ratio between sample variance and circuit variance as a natural measure of the complexity of training variational quantum circuits. We showed that careful consideration of both variances is necessary in order to determine trainability. Using this framework, we investigated the trainability of a broad class of photonic observables, identifying fixed-order photon-number polynomials as efficiently trainable: a number of circuit samples growing polynomially with system size is sufficient to train photonic circuits to minimise the expected value of such observables. We also showed that observables based on output probability estimation and high-order photon-number polynomials are not suitable for scalable variational protocols, as they have exponentially growing sample demands for loss resolution.

We further demonstrated that fixed-order photon-number polynomials admit efficient implementations through neural network observables, providing a practical and expressive class of loss functions for variational photonic protocols. We showed that these observables

appear to only require only polynomially scaling sampling resources for training and analysed the complexity of their classical simulation, identifying regimes in which quantum devices offer a polynomial advantage over existing classical methods. In Sec. 4, we validated these predictions through successful variational circuit training in both simulation and on near-term photonic hardware. Together, these results provide formal support for the empirical scalability of recent photonic quantum machine learning experiments and identify a broad family of observables for which variational photonic quantum circuits remain trainable at large system sizes.

Several theoretical questions remain open. Our analysis relies on a conjectured scaling of the gap between the first and second Haar-average moments, supported by numerical evidence but not yet established analytically. We have also restricted our treatment to fixed-photon-number input states. Extending these results to superposition states or states with indefinite photon number would provide stronger theoretical foundations for experimentally relevant regimes such as Gaussian input states. Similarly, photon losses are the dominant error channel for linear optical networks and so understanding how these errors can change the trainability landscape is important in situations where one cannot post-select on the noise-free outputs. Finally, while our variance analysis establishes when gradients can be estimated efficiently, it does not address other optimisation challenges that arise during training, such as avoiding local minima or understanding the influence of hardware imperfections.

Our results establish a theoretical foundation for understanding trainability in variational photonic quantum computing and provide practical guidance for the design of scalable photonic quantum machine learning protocols. By identifying observable classes that simultaneously avoid barren plateaus, remain sample-efficient, and admit favourable complexity compared with current classical estimation methods, this work helps bridge the gap between recent large-scale photonic demonstrations and the theoretical understanding of variational photonic quantum algorithms. We expect these results to provide a promising direction for the development of photonic quantum machine learning algorithms capable of delivering practical quantum advantage.

Acknowledgements

We thank Alex Jones, Josh Nunn and the rest of the team at ORCA Computing for insightful discussions and valuable comments on the manuscript. MSK, IAW and ZL acknowledge funding from the UK EPSRC through EP/T001062/1, EP/Z53318X/1, EP/W032643/1, and EP/T517811/1. MSK thanks the KIST through Open Innovation fund and the National Research Foundation of Korea grant funded by the Korean government (MSIT) (No. RS-2024-00413957) for their financial support. IAW is supported by the UK Research & Innovation (project number: MR/W011794/1). ZL acknowledges partial funding support from ORCA Computing.

Note added. During the preparation of this manuscript, we became aware of independent and related work by Monbroussou et al. that studies the trainability of linear optical

circuits [67].

References

1. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
2. Shor, P. W. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM review* **41**, 303–332 (1999).
3. Gidney, C. & Ekerå, M. How to factor 2048 bit RSA integers in 8 hours using 20 million noisy qubits. *Quantum* **5**, 433 (2021).
4. Cerezo, M. *et al.* Variational quantum algorithms. *Nature Reviews Physics* **3**, 625–644 (2021).
5. McClean, J. R., Romero, J., Babbush, R. & Aspuru-Guzik, A. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics* **18**, 023023 (2016).
6. Farhi, E., Goldstone, J. & Gutmann, S. A quantum approximate optimization algorithm. *arXiv preprint arXiv:1411.4028* (2014).
7. Zhou, L., Wang, S.-T., Choi, S., Pichler, H. & Lukin, M. D. Quantum approximate optimization algorithm: Performance, mechanism, and implementation on near-term devices. *Physical Review X* **10**, 021067 (2020).
8. McArdle, S., Endo, S., Aspuru-Guzik, A., Benjamin, S. C. & Yuan, X. Quantum computational chemistry. *Reviews of Modern Physics* **92**, 015003 (2020).
9. Peruzzo, A. *et al.* A variational eigenvalue solver on a photonic quantum processor. *Nature communications* **5**, 4213 (2014).
10. Biamonte, J. *et al.* Quantum machine learning. *Nature* **549**, 195–202 (2017).
11. Abbas, A. *et al.* The power of quantum neural networks. *Nature computational science* **1**, 403–409 (2021).
12. Xu, X. *et al.* Variational algorithms for linear algebra. *Science Bulletin* **66**, 2181–2188 (2021).
13. Consiglio, M. *et al.* Variational Gibbs state preparation on noisy intermediate-scale quantum devices. *Phys. Rev. A* **110**, 012445 (1 2024).
14. Zuñiga Castro, J. C. *et al.* Variational quantum state preparation for quantum-enhanced metrology in noisy systems. *Phys. Rev. A* **110**, 052615 (5 2024).
15. Jones, T., Endo, S., McArdle, S., Yuan, X. & Benjamin, S. C. Variational quantum algorithms for discovering Hamiltonian spectra. *Physical Review A* **99**, 062304 (2019).
16. Endo, S., Sun, J., Li, Y., Benjamin, S. C. & Yuan, X. Variational Quantum Simulation of General Processes. *Phys. Rev. Lett.* **125**, 010501 (1 2020).
17. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R. & Neven, H. Barren plateaus in quantum neural network training landscapes. *Nature communications* **9**, 4812 (2018).

18. Arrasmith, A., Holmes, Z., Cerezo, M. & Coles, P. J. Equivalence of quantum barren plateaus to cost concentration and narrow gorges. *Quantum Sci. Technol.* **7** (2022).
19. Larocca, M. *et al.* Barren plateaus in variational quantum computing. *Nature Reviews Physics*, 1–16 (2025).
20. Cerezo, M., Sone, A., Volkoff, T., Cincio, L. & Coles, P. J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature communications* **12**, 1791 (2021).
21. Ortiz Marrero, C., Kieferová, M. & Wiebe, N. Entanglement-Induced Barren Plateaus. *PRX Quantum* **2**, 040316 (4 2021).
22. Patti, T. L., Najafi, K., Gao, X. & Yelin, S. F. Entanglement devised barren plateau mitigation. *Physical Review Research* **3**, 033090 (2021).
23. Wang, S. *et al.* Noise-induced barren plateaus in variational quantum algorithms. *Nature communications* **12**, 6961 (2021).
24. Ragone, M. *et al.* A lie algebraic theory of barren plateaus for deep parameterized quantum circuits. *Nature Communications* **15**, 7172 (2024).
25. Larocca, M. *et al.* Diagnosing Barren Plateaus with Tools from Quantum Optimal Control. *Quantum* **6**, 824. ISSN: 2521-327X (2022).
26. Grant, E., Wossnig, L., Ostaszewski, M. & Benedetti, M. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum* **3**, 214. ISSN: 2521-327X (2019).
27. Pesah, A. *et al.* Absence of Barren Plateaus in Quantum Convolutional Neural Networks. *Phys. Rev. X* **11**, 041011 (4 2021).
28. Skolik, A., McClean, J. R., Mohseni, M., Van Der Smagt, P. & Leib, M. Layerwise learning for quantum neural networks. *Quantum Machine Intelligence* **3**, 5 (2021).
29. Sack, S. H., Medina, R. A., Michailidis, A. A., Kueng, R. & Serbyn, M. Avoiding barren plateaus using classical shadows. *PRX Quantum* **3**, 020365 (2022).
30. Cerezo, M. *et al.* Does provable absence of barren plateaus imply classical simulability? *Nature Communications* **16**, 7907 (2025).
31. Goh, M. L., Larocca, M., Cincio, L., Cerezo, M. & Sauvage, F. Lie-algebraic classical simulations for quantum computing. *Phys. Rev. Res.* **7**, 033266 (3 2025).
32. Bermejo, P. *et al.* Quantum Convolutional Neural Networks are Effectively Classically Simulable. *PRX Quantum* **7**, 020304 (2 2026).
33. Gong, S.-Q. *et al.* Enhanced Image Recognition Using Gaussian Boson Sampling. *arXiv preprint arXiv:2506.19707* (2025).
34. Cimini, V. *et al.* Large-scale quantum reservoir computing using a Gaussian Boson Sampler. *npj Quantum Information* (2026).

35. Aaronson, S. & Arkhipov, A. *The computational complexity of linear optics* in *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, San Jose, California, USA, 2011), 333–342. ISBN: 9781450306911.
36. Deshpande, A. *et al.* Quantum computational advantage via high-dimensional Gaussian boson sampling. *Science advances* **8**, eabi7894 (2022).
37. Hangleiter, D. & Eisert, J. Computational advantage of quantum random sampling. *Reviews of Modern Physics* **95**, 035001 (2023).
38. Li, Z., Solomons, N. R., Bulmer, J. F., Patel, R. B. & Walmsley, I. A. A complexity transition in displaced Gaussian Boson sampling. *npj Quantum Information* **11**, 119 (2025).
39. Li, Z. *et al.* Machine learning of quantum data using optimal similarity measurements. *arXiv preprint arXiv:2602.23501* (2026).
40. Hoch, F. *et al.* Quantum machine learning with adaptive boson sampling via post-selection. *Nature Communications* **16**, 902 (2025).
41. Kolarovszki, Z., Bakó, B., Oszmaniec, M., Oh, C. & Zimborás, Z. Generative modeling with Gaussian Boson Sampling: classically trainable Bosonic Born Machines. *arXiv preprint arXiv:2603.11195* (2026).
42. Bacarreza, O. *et al.* Quantum latent distributions in deep generative models. *arXiv preprint arXiv:2508.19857* (2025).
43. Chabaud, U., Markham, D. & Sohbi, A. Quantum machine learning with adaptive linear optics. *Quantum* **5**, 496. ISSN: 2521-327X (2021).
44. Monbroussou, L. *et al.* Toward quantum advantage with photonic state injection. *Phys. Rev. Res.* **7**, 033051 (3 2025).
45. Monbroussou, L. *et al.* Photonic quantum convolutional neural networks with adaptive state injection. *Advanced Photonics* **7**, 066012–066012 (2025).
46. Reck, M., Zeilinger, A., Bernstein, H. J. & Bertani, P. Experimental realization of any discrete unitary operator. *Phys. Rev. Lett.* **73**, 58–61 (1 1994).
47. Clements, W. R., Humphreys, P. C., Metcalf, B. J., Kolthammer, W. S. & Walmsley, I. A. Optimal design for universal multiport interferometers. *Optica* **3**, 1460–1465 (2016).
48. Bell, B. A. & Walmsley, I. A. Further compactifying linear optical unitaries. *Appl Photonics* **6** (2021).
49. Bouland, A. & Aaronson, S. Generation of universal linear optics by any beam splitter. *Physical Review A* **89**, 062316 (2014).
50. Cardin, Y. & Quesada, N. Photon-number moments and cumulants of Gaussian states. *Quantum* **8**, 1521 (2024).
51. Mhiri, H. *et al.* *Boson sampling beyond the dilute regime: second moments and anti-concentration* 2026.

52. Lim, Y. & Oh, C. *Classical Algorithms for Estimating Expectation Values in Linear-Optical Circuits* 2025.
53. Villalonga, B. *et al.* Efficient approximation of experimental Gaussian boson sampling. *arXiv preprint arXiv:2109.11525* (2021).
54. Dodd, T., Martínez-Cifuentes, J., Brown, O. T., Quesada, N. & García-Patrón, R. A fast and frugal Gaussian Boson Sampling emulator. *arXiv preprint arXiv:2511.14923* (2025).
55. Recio-Armengol, E., Ahmed, S. & Bowles, J. Train on classical, deploy on quantum: scaling generative quantum machine learning to a thousand qubits. *arXiv preprint arXiv:2503.02934* (2025).
56. Brádler, K., Friedland, S., Izaac, J., Killoran, N. & Su, D. Graph isomorphism and Gaussian boson sampling. *arXiv preprint arXiv:1810.10644* (2018).
57. Mezher, R., Carvalho, A. F. & Mansfield, S. Solving graph problems with single photons and linear optics. *Physical Review A* **108**, 032405 (2023).
58. Nikolopoulos, G. M. & Brougham, T. Decision and function problems based on boson sampling. *Physical Review A* **94**, 012315 (2016).
59. Wang, S.-T. & Duan, L.-M. Certification of boson sampling devices with coarse-grained measurements. *arXiv preprint arXiv:1601.02627* (2016).
60. Singh, D. *et al.* Proof-of-work consensus by quantum sampling. *Quantum Science and Technology* **10**, 025020 (2025).
61. Nikolopoulos, G. M. Cryptographic one-way function based on boson sampling: GM Nikolopoulos. *Quantum Information Processing* **18**, 259 (2019).
62. Novák, S., Roberts, D. D., Makarovskiy, A., García-Patrón, R. & Clements, W. R. Boundaries for quantum advantage with single photons and loop-based time-bin interferometers. *Quantum* **9**, 1915 (2025).
63. Makarovskiy, A. *et al.* A Binary Optimisation Algorithm for Near-Term Photonic Quantum Processors. *arXiv preprint arXiv:2510.08274* (2025).
64. Cazalis, J., Chai, Y., Jansen, K., Kühn, S. & Shah, T. *Gaussian Boson Sampling for binary optimization* in *2023 IEEE International Conference on Quantum Computing and Engineering (QCE)* **2** (2023), 332–333.
65. Lee, D. *et al.* Error-mitigated photonic variational quantum eigensolver using a single-photon ququart. *Optica* **9**, 88–95 (2022).
66. Taylor, A., Bressanini, G., Kwon, H. & Kim, M. S. Quantum error cancellation in photonic systems: Undoing photon losses. *Phys. Rev. A* **110**, 022622 (2 2024).
67. Monbroussou, L., Thomas, H., Mhiri, H., Holmes, Z. & Kashefi, E. *Classical simulation and model concentration in passive linear optics* 2026.
68. Facelli, G. *et al.* Exact gradients for linear optics with single photons. *arXiv preprint arXiv:2409.16369* (2024).

69. Pappalardo, A. *et al.* Photonic parameter-shift rule: Enabling gradient computation for photonic quantum computers. *Physical Review A* **111**, 032429 (2025).
70. Russell, N. J., Chakhmakhchyan, L., O’Brien, J. L. & Laing, A. Direct dialling of Haar random unitary matrices. *New journal of physics* **19**, 033007 (2017).
71. Efron, B. & Stein, C. The Jackknife Estimate of Variance. *The Annals of Statistics* **9**, 586–596 (1981).
72. Isoperimetric and analytic inequalities for log-concave probability measures. *The Annals of Probability* **27**, 1903–1921 (1999).
73. Collins, B. & Śniady, P. Integration with respect to the Haar measure on unitary, orthogonal and symplectic group. *Communications in Mathematical Physics* **264**, 773–795 (2006).
74. Collins, B., Matsumoto, S. & Saad, N. Integration of invariant matrices and moments of inverses of Ginibre and Wishart matrices. *Journal of Multivariate Analysis* **126**, 1–13 (2014).
75. Taylor, A., Hanks, M., Kwon, H. & Kim, M. S. *Optimal Quantum Information Transmission Under a Continuous-Variable Erasure Channel* 2025.
76. Belovas, I. *Mathematica Slovaca* **73**, 335–342 (2023).
77. Trudeau, M., Emeriau, P.-E. & Quesada, N. Parameter-Shift Rules for Gradients in Boson Sampling Experiments. *arXiv preprint arXiv:2607.15160* (2026).
78. Hoch, F. *et al.* Variational approach to photonic quantum circuits via the parameter shift rule. *Physical Review Research* **7**, 023227 (2025).
79. Wierichs, D., Izaac, J., Wang, C. & Lin, C. Y.-Y. General parameter-shift rules for quantum gradients. *Quantum* **6**, 677 (2022).

A Sample complexity for resolving loss differences

In this section we justify the claims of Sec. 2.2 that a number of samples scaling with the variance ratio $\text{Var}_{\text{samp}}(\hat{O})/\text{Var}_{\text{circ}}(\hat{O})$ is sufficient to resolve loss differences at different parameter settings to within proportional error. We do this through Chebyshev’s inequality: setting the precision of the loss estimates at the standard deviation of the loss landscape, a number of samples scaling with the variance ratio suffices to resolve any pair of parameter settings whose losses differ at this scale. Under weak conditions on the loss distribution, a non-negligible fraction of randomly drawn pairs are guaranteed to differ on this scale, justifying the use of this ratio as a measure of trainability.

A.1 Setup

We consider a fixed parameter setting $\boldsymbol{\theta}$. A single measurement of the observable $\hat{O}$ on $\rho(\boldsymbol{\theta})$ returns an outcome $\hat{O}^{(i)}$ drawn from a distribution with mean $\ell(\boldsymbol{\theta}) = \text{tr}[\hat{O}\rho(\boldsymbol{\theta})]$ and variance

$$\text{Var}[\hat{O}|\boldsymbol{\theta}] = \text{tr}[\hat{O}^2\rho(\boldsymbol{\theta})] - \ell(\boldsymbol{\theta})^2. \quad (27)$$

Given N independent measurement outcomes $\hat{O}^{(1)}, \dots, \hat{O}^{(N)}$, we form the standard empirical loss estimate:

$$\bar{\ell}_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^N \hat{O}^{(i)}, \quad (28)$$

which satisfies $\mathbb{E}[\bar{\ell}_N(\boldsymbol{\theta})] = \ell(\boldsymbol{\theta})$ and, by independence of the $\hat{O}^{(i)}$ measurements,

$$\text{Var}[\bar{\ell}_N(\boldsymbol{\theta})] = \frac{\text{Var}[\hat{O}|\boldsymbol{\theta}]}{N}. \quad (29)$$

A.2 Sampling requirements through Chebyshev’s inequality

Directly applying Chebyshev’s inequality, we find that for any $\varepsilon_{\text{additive}} > 0$,

$$\Pr \left[|\bar{\ell}_N(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})| \geq \varepsilon_{\text{additive}} \right] \leq \frac{\text{Var}[\hat{O}|\boldsymbol{\theta}]}{N\varepsilon_{\text{additive}}^2}. \quad (30)$$

In particular, for any $\delta \in (0, 1]$, taking $N \geq \text{Var}[\hat{O}|\boldsymbol{\theta}]/(\delta\varepsilon_{\text{additive}}^2)$ samples suffices to estimate $\ell(\boldsymbol{\theta})$ to within additive error $\varepsilon_{\text{additive}}$ with probability at least $1 - \delta$.

In order to resolve a loss-decreasing direction with parameter shift rules, it is necessary to resolve which of two parameter settings, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}'$, has the lowest loss. Although photonic parameter-shift rules may generally involve several loss evaluations, we focus on this two-point comparison as a useful case study. We expect the resulting conclusions to extend naturally to more general parameter-shift rules.

We’ll denote the difference in the losses by $D(\boldsymbol{\theta}, \boldsymbol{\theta}') = \ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}')$, and the estimated difference in losses by $\bar{D}_N(\boldsymbol{\theta}, \boldsymbol{\theta}') = \bar{\ell}_N(\boldsymbol{\theta}) - \bar{\ell}_N(\boldsymbol{\theta}')$. We have $\mathbb{E}[\bar{D}_N] = D$ and, by indepen-

dence of the observable samples,

$$\text{Var}[\bar{D}_N] = \frac{\text{Var}[\hat{O}|\boldsymbol{\theta}] + \text{Var}[\hat{O}|\boldsymbol{\theta}']}{N}. \quad (31)$$

We take the magnitude of the loss difference, $|D(\boldsymbol{\theta}, \boldsymbol{\theta}')|$, to define the relevant error scale. Estimating $\bar{D}_N$ to within a relative error $\varepsilon_{\text{relative}}|D|$ is sufficient to determine the loss decreasing direction and, in the context of parameter shift rules, estimate the gradient with a relative accuracy scaling as $\varepsilon_{\text{relative}}$. Plugging this error into the above inequality, we know that in order to achieve this with constant probability $1 - \delta$, we see the number of shots is bounded by

$$N \geq \frac{\text{Var}[\hat{O}|\boldsymbol{\theta}] + \text{Var}[\hat{O}|\boldsymbol{\theta}']}{\delta \varepsilon^2 (\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}'))^2} \quad (32)$$

A.3 Sufficiency of the variance ratio

Equation (32) bounds the samples needed to resolve the loss at a pair of parameter settings in terms of the fixed-parameter variances at the two settings and their loss difference D .

When $\boldsymbol{\theta}, \boldsymbol{\theta}'$ are drawn independently, averaging over both parameters recovers the sample variance and circuit variance discussed in Sec. 2. Explicitly, the numerator averages to $\mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\theta}'} [\text{Var}[\hat{O}|\boldsymbol{\theta}] + \text{Var}[\hat{O}|\boldsymbol{\theta}']] = 2 \text{Var}_{\text{samp}}(\hat{O})$. Similarly, because $\mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\theta}'} [D(\boldsymbol{\theta}, \boldsymbol{\theta}')] = 0$, the expected value of the squared loss difference in the denominator is exactly its variance over the landscape,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}, \boldsymbol{\theta}'} [D^2(\boldsymbol{\theta}, \boldsymbol{\theta}')] &= \text{Var}_{\boldsymbol{\theta}, \boldsymbol{\theta}'} [D(\boldsymbol{\theta}, \boldsymbol{\theta}')] \\ &= \text{Var}_{\boldsymbol{\theta}} [\ell(\boldsymbol{\theta})] + \text{Var}_{\boldsymbol{\theta}'} [\ell(\boldsymbol{\theta}')] \\ &= 2 \text{Var}_{\text{circ}}(\hat{O}), \end{aligned} \quad (33)$$

using independence of the parameter choices. At the level of these averages, the pairwise bound (32) therefore becomes exactly the variance ratio, $N \geq \text{Var}_{\text{samp}}(\hat{O}) / (\delta \varepsilon^2 \text{Var}_{\text{circ}}(\hat{O}))$.

In particular, Eq. (33) identifies $\sqrt{\text{Var}_{\text{circ}}(\hat{O})}$ as the characteristic scale of loss differences across the circuit ensemble. It is therefore natural to define the target additive error as a fixed fraction of this scale, $\varepsilon_{\text{additive}} = \varepsilon \sqrt{\text{Var}_{\text{circ}}(\hat{O})}$ so ε can be interpreted as the relative error. Estimating each loss to this accuracy corresponds to resolving variations that are proportional to typical differences encountered across the loss landscape; since $\sqrt{\text{Var}_{\text{circ}}(\hat{O})}$ quantifies the characteristic magnitude of these fluctuations, this choice of error scale should allow changes induced by varying the circuit parameters to be distinguished. Substituting this error scale into Chebyshev's inequality then naturally gives rise to the ratio between the expected sample variance and the circuit variance.

We will now formalise this statement.

Proposition 1. *Let $\varepsilon > 0$, $\delta \in (0, 1]$ and $t > 0$. Let $\boldsymbol{\theta}, \boldsymbol{\theta}'$ be drawn independently from a*

parameter distribution. With probability at least $1 - 2/t$, taking

$$N \geq \frac{t}{\delta \varepsilon^2} \frac{\text{Var}_{\text{samp}}(\hat{O})}{\text{Var}_{\text{circ}}(\hat{O})} \quad (34)$$

samples at each setting yields loss estimates satisfying

$$|\bar{\ell}_N(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})|, |\bar{\ell}_N(\boldsymbol{\theta}') - \ell(\boldsymbol{\theta}')| \leq \varepsilon \sqrt{\text{Var}_{\text{circ}}(\hat{O})}, \quad (35)$$

simultaneously with probability at least $1 - 2\delta$. In particular, whenever $|\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}')| > 2\varepsilon \sqrt{\text{Var}_{\text{circ}}(\hat{O})}$, the loss-minimising direction is correctly identified.

Proof. The sample variance is the average of the single-shot variances, $\text{Var}_{\text{samp}}(\hat{O}) = \mathbb{E}_{\boldsymbol{\theta}} [\text{Var}[\hat{O}|\boldsymbol{\theta}]]$, so by Markov's inequality $\mathbb{P}_{\boldsymbol{\theta}} [\text{Var}[\hat{O}|\boldsymbol{\theta}] > t \text{Var}_{\text{samp}}(\hat{O})] < 1/t$, giving $\text{Var}[\hat{O}|\boldsymbol{\theta}], \text{Var}[\hat{O}|\boldsymbol{\theta}'] \leq t \text{Var}_{\text{samp}}(\hat{O})$ with probability at least $1 - 2/t$. On this event, Chebyshev's inequality (30) with error $\varepsilon_{\text{additive}} = \varepsilon \sqrt{\text{Var}_{\text{circ}}(\hat{O})}$ shows that each estimate achieves this error with probability at least $1 - \delta$ for N satisfying (34). Both estimates achieve these errors simultaneously with probability at least $1 - 2\delta$. Finally, if $|D| > 2\varepsilon \sqrt{\text{Var}_{\text{circ}}(\hat{O})}$, the triangle inequality gives $\text{sign}(\bar{D}_N) = \text{sign}(D)$. $\square$

Taking t, ε and δ to be constants, a number of samples $N = \mathcal{O}(\text{Var}_{\text{samp}}(\hat{O})/\text{Var}_{\text{circ}}(\hat{O}))$ is sufficient to resolve losses to within their expected fluctuation, and to correctly order any pair of settings whose losses differ by more than this resolution.

Proposition 1 makes explicit how trainability fails at the two ends of the variance ratio. If the sample variance is exponentially large in system size, it takes exponentially many samples to build estimates to fixed precision. If instead the circuit variance is exponentially small, the loss differences need to be resolved to exponentially small error $\sqrt{\text{Var}_{\text{circ}}(\hat{O})}$, and by (30) reaching this precision requires exponentially many samples for constant or larger shot noise. In either case the sample complexity (34) is exponential unless the other variance compensates. Conversely, if the sample variance grows at most polynomially with system size and the circuit variance decays at most polynomially, then the variance ratio is at most polynomial, and polynomially many samples are certainly sufficient to resolve loss differences at the landscape scale.

A.4 Typical loss differences

With proposition 1, we show that with a number of samples scaling with the variance ratio, we can resolve the difference between loss at two different parameter settings to within the average loss fluctuations expected in the landscape. However, we don't consider how regularly we can expect the difference of loss to be of this magnitude. We now discuss general assumptions on the loss landscape that would be sufficient to guarantee that values of $|D|$ scale approximately with $\sqrt{\text{Var}_{\text{circ}}(\hat{O})}$ with high probability.

(i) *Anticoncentration of the landscape.* There exist $\alpha, \beta > 0$, scaling at worst inverse-

polynomially in system size, such that

$$\mathbb{P}\left[D^2 \geq \beta \cdot 2 \text{Var}_{\text{circ}}(\hat{O})\right] \geq \alpha. \quad (36)$$

In words, this conjectures that for an inverse-polynomially large fraction of randomly drawn pairs, the loss difference is at least as large as $\sqrt{\text{Var}_{\text{circ}}(\hat{O})}$ up to a polynomial factor. Under this condition, proposition 1 with any $\varepsilon < \sqrt{\beta/2}$, using $N = O(\beta^{-1} \text{Var}_{\text{samp}}(\hat{O})/\text{Var}_{\text{circ}}(\hat{O}))$ samples is sufficient for loss resolution. As such, having a polynomially scaling variance ratio is still sufficient to claim efficient trainability, by our definition. We note that this condition is analogous to the anticoncentration properties of boson sampling output probabilities [51]. Indeed this condition holds for $\hat{O} = |\mathbf{s}\rangle\langle\mathbf{s}|$ estimating output probabilities through the anticoncentration conjecture.

- (ii) *Gaussian loss differences.* We assume $D \sim \mathcal{N}(0, 2 \text{Var}_{\text{circ}}(\hat{O}))$. This gives condition (i) with sharp, system-size-independent constants:

$$\mathbb{P}\left[|D| \geq \sqrt{\text{Var}_{\text{circ}}(\hat{O})}\right] = 2 \Phi(-1/\sqrt{2}) \approx 0.48, \quad (37)$$

where Φ is the standard normal cumulative distribution function. We can conclude nearly half of all randomly drawn pairs are resolved by proposition 1 with any $\varepsilon < 1/2$ and $N = O(\text{Var}_{\text{samp}}(\hat{O})/\text{Var}_{\text{circ}}(\hat{O}))$ samples, with no polynomial overheads.

Under either condition, combined with proposition 1, a large fraction of parameter pairs are resolvable with a number of samples scaling, up to polynomial factors, as the variance ratio $\text{Var}_{\text{samp}}(\hat{O})/\text{Var}_{\text{circ}}(\hat{O})$, which is then sufficient as a measure of trainability. Together with the equivalence between gradient and loss concentration established in Appendix B, this extends from resolving loss differences to resolving gradients themselves.

B Equivalence between barren plateaus and loss concentration

In Ref. [18], the authors prove an equivalence between exponential concentration of the loss function and the presence of barren plateaus across all parameters. This implies that loss concentration is a sufficient condition for diagnosing barren plateaus, but the converse is weaker; a non-concentrated loss does not guarantee that every parameter remains trainable, but only that at least one parameter has a non-vanishing gradient. To determine the absence of barren plateaus across all parameters requires additionally showing that the parameters variances are in the same trainability class, for example by showing that all gradient variances are comparable in size.

In this section, we first adapt the results of Ref. [18] to naturally parametrised linear optical photonic architectures. A key difference that emerges is that average gradients can be non-zero, and hence concentrated gradient variance alone is insufficient to determine whether a parameter is trainable. We therefore look at the second moment of the gradients

as the analogous condition for whether a parameter is trainable. With this, we show an equivalence between loss concentration and all parameter admitting barren plateau type behaviour.

We then consider a particular class of parametrised linear optical networks and show that the ensemble averaged gradient second moments are equal for all parameters. For this class of circuits, we can expect that anti-concentration of the loss function implies the absence of barren plateaus across all parameters for this class of circuits.

B.1 Setup

Consider an M -mode parametrised linear optical network consisting of L parameters with associated generator $\{\hat{B}_j\}$. This can be written as

$$U(\boldsymbol{\theta}) = \prod_{j=1}^L U_j(\theta_j), \quad U_j(\theta_j) = e^{-i\theta_j \hat{B}_j} \hat{W}_j \quad (38)$$

where $\hat{W}_j$ is a fixed unitary. For example, the generators could be beam splitter interaction and fixed unitaries fixed phase shifters. The gradient of this parametrised circuit with respect to θ_k reads

$$G_k(\boldsymbol{\theta}) = U_{+k}(\boldsymbol{\theta}) (-i\hat{B}_k) U_{-k}(\boldsymbol{\theta}) \quad (39)$$

where we defined

$$U_{-k}(\boldsymbol{\theta}) = \prod_{j=1}^k U_j(\boldsymbol{\theta}), \quad U_{+k}(\boldsymbol{\theta}) = \prod_{j=k+1}^L U_j(\boldsymbol{\theta}) \quad (40)$$

satisfying $U_{+j}U_{-j} = U$.

Now we consider a loss function $\ell(\boldsymbol{\theta})$ given by the expectation value of some observable

$$\ell(\boldsymbol{\theta}) = \text{tr} \left[U(\boldsymbol{\theta}) \rho U^\dagger(\boldsymbol{\theta}) \hat{O} \right] \quad (41)$$

The squared magnitude of loss fluctuation across the parameter landscape is captured by the variance

$$\text{Var}_{\boldsymbol{\theta}}[\ell(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[\ell(\boldsymbol{\theta})^2] - \mathbb{E}_{\boldsymbol{\theta}}[\ell(\boldsymbol{\theta})]^2 \quad (42)$$

A loss function is said to be concentrated if the size of fluctuations is exponentially small with respect to problem size N

$$\text{Var}_{\boldsymbol{\theta}}[\ell(\boldsymbol{\theta})] \sim O(e^{-N}) \implies \ell \text{ is exponentially concentrated.} \quad (43)$$

In qubit-based architectures, the problem size is almost always determined by the qubit count. In our photonic case, there is more freedom how problem size is defined – the number of modes, number of input photons, or support/weight of the observable can all

meaningfully change the problem size. For now, we keep the problem size as labelled by N .

The derivative of the loss function with respect to a parameter θ_k , denoted $G_k(\boldsymbol{\theta})$, can be written as

$$\begin{aligned} G_k(\boldsymbol{\theta}) &= \partial_{\theta_k} \ell(\boldsymbol{\theta}) \\ &= i \operatorname{tr} [U_{-k}(\boldsymbol{\theta}) \rho U_{-k}^\dagger(\boldsymbol{\theta}) [\hat{B}_k, U_{+k}^\dagger(\boldsymbol{\theta}) \hat{O} U_{+k}(\boldsymbol{\theta})]] \end{aligned} \quad (44)$$

where $[\bullet, \star] = \bullet \star - \star \bullet$ is the commutator. In qubit devices, the average gradient is zero [17–19]. In photonic devices, if averages are taken with respect to the Haar measure then this is not the case because $\mu_{\text{Haar}}(\boldsymbol{\theta}) \neq \text{Uniform}([0, 2\pi])$; instead there can be a specific generator-observable dependence,

$$\mathbb{E}_{\boldsymbol{\theta}}[G_k(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}} \left[i \operatorname{tr} \left[U_{-k}(\boldsymbol{\theta}) \rho U_{-k}^\dagger(\boldsymbol{\theta}) [\hat{B}_k, U_{+k}^\dagger(\boldsymbol{\theta}) \hat{O} U_{+k}(\boldsymbol{\theta})] \right] \right]. \quad (45)$$

Barren plateaus are said to occur if the gradient variances decrease exponentially with respect to the problem size N [17–19]

$$\operatorname{Var}_{\boldsymbol{\theta}}[G_k(\boldsymbol{\theta})] \sim O(e^{-N}) \implies \ell \text{ has a } \theta_k \text{ barren plateau.} \quad (46)$$

A barren plateau so the k^{th} parameter with average gradient $\mathbb{E}_{\boldsymbol{\theta}}[G_k(\boldsymbol{\theta})] = 0$ indicates that exponential precision is required in order to distinguish the direction (positive or negative) in which the loss function is decreasing. Therefore, optimising the loss function along the θ_k manifold typically requires exponentially many shots and is not efficient. In contrast, if the average gradient is non-zero then estimating the sign of the gradient is far easier; if one can efficiently estimate if G_k is positive or negative then one can update parameters in the correct direction efficiently. This suggests that the variance-based barren plateau definition is not suitable for cases when $\mathbb{E}_{\boldsymbol{\theta}}[G_k]$ can be non-zero. For example, we could consider gradients being exponentially concentrated around some polynomially scaling average $|\mathbb{E}_{\boldsymbol{\theta}}[G_k]| > O(e^{-N})$; while the variance of G_k indicates the existence of barren plateaus, the parameter θ_k could actually be efficiently updated.

In the case with potentially finite average gradient, it is more suitable to define barren plateaus with respect to the expected gradient second moments; we say the k^{th} parameter admits a barren plateau if

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[G_k^2(\boldsymbol{\theta})] &= \operatorname{Var}_{\boldsymbol{\theta}}[G_k(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}}[G_k(\boldsymbol{\theta})]^2 \\ &\sim O(e^{-N}) \implies \ell \text{ has a } \theta_k \text{ barren plateau.} \end{aligned} \quad (47)$$

This now includes a term accounting for the squared magnitude of the average gradient. If the average gradient is zero, we recover the standard definition. Similarly, if both the average and variance terms are exponentially small then this correctly indicates the parameter is not efficiently trainable. However, if the variance is exponentially small but the average is polynomially scaling, then the gradient second moment is also polynomially scaling and our definition indicates that the parameter is trainable.

For this new definition of barren plateau, we go on to prove a photonic analogue of the results from Ref. [18]; that there is an equivalence between the loss function being exponentially concentrated and all parameters admitting barren plateaus. First, we show that loss concentration implies barren plateaus.

B.2 Loss concentration implies barren plateaus

In Ref. [68] and Ref. [69], the authors prove a parameter shift rule for linear optical networks with Fock input states. Namely, for a loss function $\ell(\boldsymbol{\theta}) = \text{tr}[U(\boldsymbol{\theta})\rho U^\dagger(\boldsymbol{\theta})\hat{O}]$ and input Fock state containing a definite n photons, the gradients are given by [68]

$$\partial_{\theta_k} \ell(\boldsymbol{\theta}) = \sum_{j=0}^{2n} \frac{(-1)^{j+1}}{4n \sin^2(\kappa_j/2)} \ell(\boldsymbol{\theta} + \kappa_j \mathbf{e}_k) \quad (48)$$

for $\kappa_j = \frac{(2j-1)\pi}{2n}$. For observables given by photon number monomials (PNMs) of degree $|\boldsymbol{\alpha}|$, the number of loss evaluations can be reduced to $\min\{|\boldsymbol{\alpha}|, n\}$ [68]. This implies that for observables with a degree scaling independently of system size, the number of loss evaluations is constant and given by $|\boldsymbol{\alpha}|$.

Loss concentration implying barren plateaus is a direct consequence of the parameter shift rule, as proven in Ref. [18]. The standard qubit-based derivation has a parameter shift rule containing two terms. In our photonic setup, the gradient can be written as a sum over $\min\{|\boldsymbol{\alpha}|, n\}$ loss evaluations. Using the same argument as in Ref. [18], it is clear that a concentrated cost function implies all parameters admit barren plateaus, because the sum of polynomially many exponentially concentrated loss functions must itself be exponentially concentrated. This proves the following statement;

Proposition 2. *If a loss function $\ell(\boldsymbol{\theta})$ is exponentially concentrated, then this implies that all associated gradients $G_k = \partial_{\theta_k} \ell(\boldsymbol{\theta})$ have an exponentially concentrated second moment and hence admit barren plateaus.*

B.3 Barren plateaus imply loss concentration

In Ref. [18], the derivation that barren plateaus imply cost concentration applied if (i) the loss function was periodic on each parameter and (ii) the parameters were sampled from a uniform distribution on their domain. In the photonic case, we are interested in sampling from Haar random linear optical networks for which both (i) and (ii) are not simultaneously true. For example, consider using the method of direct-dialling Haar random linear optical networks in Ref. [70] – the beam splitter parameters are not uniformly sampled and hence we need a different method of showing barren plateaus imply loss concentration.

Each beam splitter is implemented as a Mach–Zehnder interferometer with both an internal and external phase, and the circuit ends with a layer of M output phases, giving $L = M^2$ trainable parameters. Russel *et al* [70] construct a product measure $\nu = \nu_1 \times \dots \times \nu_L$ on $\boldsymbol{\theta}$ such that $U(\boldsymbol{\theta}) \sim \text{Haar}(U(M))$ whenever $\boldsymbol{\theta} \sim \nu$. The external output phases are uniform on $[0, 2\pi)$ while the internal phases at layer $\in \{1, \dots, M-1\}$ have probability

density

$$p_k(\theta) = k \sin^{2k-1}(\theta/2) \cos(\theta/2) \quad (49)$$

for $\theta \in [0, \pi]$. We call ν the Haar-dialling measure; it satisfies $\text{Var}_{\text{circ}}(\hat{O}) = \text{Var}_{\boldsymbol{\theta}}[\ell(\boldsymbol{\theta})]$ being exactly the circuit variance from Sec. 3.

We now prove the following statement, that the circuit variance is upper bounded by the largest gradient second moment.

Theorem 2. *For $\boldsymbol{\theta}$ drawn from the Haar-dialling measure ν of Ref. [70], the circuit variance of the loss function is upper bounded by*

$$\begin{aligned} \text{Var}_{\text{circ}}(\hat{O}) &\leq \sum_{j=1}^L \kappa_j \mathbb{E}_{\boldsymbol{\theta} \sim \nu} [G_j(\boldsymbol{\theta})^2] \\ &\leq \frac{3L\pi^2}{2} \max_j \mathbb{E}_{\boldsymbol{\theta} \sim \nu} [G_j(\boldsymbol{\theta})^2] \end{aligned} \quad (50)$$

where $\kappa_j = 1$ for external and output phases, and $\kappa_j \leq 3\pi^2/2$ for internal phases.

Proof. The random variables θ_j are independent under this measure ν , so the Efron-Stein inequality [71] gives

$$\text{Var}_{\boldsymbol{\theta} \sim \nu}[\ell(\boldsymbol{\theta})] \leq \sum_{j=1}^L \mathbb{E}_{\boldsymbol{\theta} \sim \nu} [\text{Var}_{\theta_j}[\ell(\boldsymbol{\theta})|\boldsymbol{\theta}_{\setminus j}]] \quad (51)$$

where $\text{Var}_{\theta_j \sim \nu_j}[\bullet|\boldsymbol{\theta}_{\setminus j}]$ is the variance over $\theta_j \sim \nu_j$ with all other coordinates held fixed. Each conditional variance can then be bounded by a one-dimensional Poincaré inequality of the form $\text{Var}_{\theta_j \sim \nu_j}[f(\theta_j)] \leq \kappa_j \mathbb{E}_{\theta_j \sim \nu_j}[(\partial_{\theta_j} f(\theta_j))^2]$.

For the uniformly-distributed external and output phases, Wirtinger’s inequality gives $\kappa_j = 1$. For the internal phases, the probability density (49) is log-concave on $(0, \pi)$, since

$$\frac{d^2}{d\theta^2} \log p_k(\theta) = -\frac{2k-1}{4} \csc^2(\theta/2) - \frac{1}{4} \sec^2(\theta/2) < 0, \quad (52)$$

Thus, Bobkov’s Poincaré inequality for log-concave measures on $\mathbb{R}$ [72] gives $\kappa_j \leq 12 \text{Var}_{\nu_j}[\theta_j]$; since $\theta_j \in [0, \pi]$, Popoviciu’s inequality bounds $\text{Var}_{\nu_j}[\theta_j] \leq \pi^2/4$, so $\kappa_j \leq 3\pi^2/2$.

Substituting these Poincaré inequalities into (51) and taking expectation over the remaining coordinates gives the first inequality in (50), where we’ve bounded every κ_j by $3\pi^2/2$. $\square$

Theorem 2 states that if all parameters admit barren plateaus (as indicated by an exponentially small $\mathbb{E}_{\boldsymbol{\theta}}[G_k]$), then the loss function is also exponentially concentrated. This therefore proves that barren plateaus imply loss concentration.

Together with the last section, these results demonstrate an equivalence between exponential concentration of the loss function and the presence of barren plateaus across all circuit parameters. We next identify a class of parametrised circuits for which the ensemble-average gradient variances are equal for every parameter. The variance of the

loss function can therefore be expected to provide a necessary and sufficient condition for understanding whether barren plateaus are present or not across the entire parameter landscape.

B.4 Equality of expected gradient second moments

A polynomially scaling loss function variance alone is insufficient to rule out barren plateaus – it merely indicates that *at least one parameter* is trainable, but perhaps a further $L - 1$ parameters are untrainable. In order to rule out barren plateaus by considering only the loss variance, we will now consider a family of parametrised circuits for which the expected value of the gradient second moments, $\mathbb{E}[G_k^2]$, are all equal. In this case, the above equivalence between barren plateaus and concentrated loss function implies that a polynomially scaling loss variance is indeed sufficient to determine that all parameters are trainable.

However, for a generic parametrised circuit, we can never expect all gradient variances to be in the same trainability class because some parameters may inevitably have zero impact on the loss function. Take for example an initial state $|1, 0, \dots, 0\rangle$; clearly any beam splitter outside the lightcone of the first mode will have zero impact on the loss function and hence the associated parameter gradients are exactly zero. A similar argument applies to local observables.

In this section, we consider sandwiching the parametrised network $U(\boldsymbol{\theta}) = \prod_{j=1}^L U_j(\theta_j)$ between two other linear optical networks V_1 and V_2 drawn from the Haar measure of $U(M)$ such that the total dynamics read

$$U_{\text{total}}(\boldsymbol{\theta}; V_1, V_2) = V_2 U(\boldsymbol{\theta}) V_1. \quad (53)$$

These networks distribute the state and observable across all M modes and can be expected to prevent gradients from being identically zero. The loss function and its derivatives can be written as

$$\ell(\boldsymbol{\theta}; V_1, V_2) = \text{tr} \left[U(\boldsymbol{\theta}) V_1 \rho V_1^\dagger U^\dagger(\boldsymbol{\theta}) V_2^\dagger \hat{O} V_2 \right] \quad (54)$$

and

$$\begin{aligned} G_k(\boldsymbol{\theta}; V_1, V_2) &= \partial_{\theta_k} \ell(\boldsymbol{\theta}; V_1, V_2) \\ &= i \text{tr} \left[U_{-k}(\boldsymbol{\theta}) V_1 \rho V_1^\dagger U_{-k}(\boldsymbol{\theta}) \left[\hat{B}_k, U_{+k}^\dagger(\boldsymbol{\theta}) V_2^\dagger \hat{O} V_2 U_{+k}(\boldsymbol{\theta}) \right] \right] \end{aligned} \quad (55)$$

respectively. As mentioned, the average gradient values can be non-zero and so it is the second moment of the gradient,

$$\begin{aligned} \mu_2(G_k) &= \mathbb{E}_{\boldsymbol{\theta}} [G_k(\boldsymbol{\theta})^2] \\ &= \text{Var}_{\boldsymbol{\theta}} [G_k(\boldsymbol{\theta})] + \mathbb{E}_{\boldsymbol{\theta}} [G_k(\boldsymbol{\theta})]^2 \end{aligned} \quad (56)$$

that can be used to identify barren plateau behaviour – if this quantity is not exponentially

small, then one does not require exponential precision in order to update the parameter θ_k in the correct direction. We now state and prove the following proposition.

Proposition 3. *Let $G_j(\boldsymbol{\theta}; V_1, V_2)$ and $G_k(\boldsymbol{\theta}; V_1, V_2)$ denote the gradients associated with parameters θ_j and θ_k respectively, which have generators $\hat{B}_j$ and $\hat{B}_k$. If the two generators are equivalent up to a permutation of the mode labels, then their expected gradient second moments are equal. More precisely, if*

$$\hat{B}_j = \hat{\pi} \circ \hat{B}_k \quad (57)$$

for some operator $\hat{\pi}$ that permutes the mode labels, then

$$\mathbb{E}_{V_1, V_2} [\mu_2(G_j(V_1, V_2))] = \mathbb{E}_{V_1, V_2} [\mu_2(G_k(V_1, V_2))]. \quad (58)$$

Consequently, all parameters whose generators belong to the same equivalence class under mode permutations have the same trainability scaling. In particular, when averaging over V_1 and V_2 , all beam-splitter parameters either admit barren-plateau-type behaviour or do not.

Proof. We can directly compute the ensemble averaged value of the second moment over the Haar random networks V_1 and V_2 . The second moment of the gradient G_j over $\boldsymbol{\theta}$ can be written as

$$\mu_2(G_k(V_1, V_2)) = -\mathbb{E}_{\boldsymbol{\theta}} \text{tr} \left[(U_{-k} V_1 \rho V_1^\dagger U_{-k}^\dagger)^{\otimes 2} [\hat{B}_k, U_{+k}^\dagger V_2^\dagger \hat{O} V_2 U_{+k}]^{\otimes 2} \right] \quad (59)$$

where we dropped the $\boldsymbol{\theta}$ argument on $U_{\pm k}$ for ease of reading. Taking the average over V_1 and V_2 gives

$$\mathbb{E}_{V_1, V_2} [\mu_2(G_k(V_1, V_2))] = -\mathbb{E}_{V_1, V_2, \boldsymbol{\theta}} \text{tr} \left[(U_{-k} V_1 \rho V_1^\dagger U_{-k}^\dagger)^{\otimes 2} [\hat{B}_k, U_{+k}^\dagger V_2^\dagger \hat{O} V_2 U_{+k}]^{\otimes 2} \right] \quad (60)$$

Using the left/right invariance of the Haar measure, we make the transformation $V_1 \rightarrow U_{-k}^\dagger V_1$ and $V_2 \rightarrow V_2 U_{+k}^\dagger$ to remove all U dependence.

$$\mathbb{E}_{V_1, V_2} [\mu_2(G_k(V_1, V_2))] = -\text{tr} \left[\mathbb{E}_{V_1} [(V_1 \rho V_1^\dagger)^{\otimes 2}] \mathbb{E}_{V_2} [\hat{B}_k, V_2 \hat{O} V_2^\dagger]^{\otimes 2} \right] \quad (61)$$

Now we notice that, due to the symmetry of Haar averaged operators, this quantity depends only on what class of generators $\hat{B}_k$ is in; for our linear optical network, this is either a beam splitter or phase shifter. Therefore, the ensemble averaged gradient second moments for all parameters that share a generator are equal, proving the above proposition. $\square$

Finally, consider parameterised circuits containing trainable beam splitters and randomly chosen, fixed phase shifters. Let $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ denote the beam-splitter and phase-shifter parameters, respectively, and suppose that only $\boldsymbol{\theta}$ is varied during training. Since the Haar measure induced by the parameterisation of Ref. [70] factorises over $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$, the two sets

of parameters may be averaged independently. We may therefore treat the average over the fixed phase shifters as an ensemble average and study the expected beam-splitter gradient second moments. This avoids the possibility that a non-vanishing loss variance is supported entirely by trainable phase-shifter directions while all beam-splitter gradients remain exponentially suppressed. We find that, after averaging over the random phase shifters and the Haar-random unitaries V_1 and V_2 , the beam-splitter gradient second moments are equal for all trainable parameters – for $G_j(\boldsymbol{\theta}; \boldsymbol{\varphi}, V_1, V_2) = \partial_{\theta_j} \ell(\boldsymbol{\theta}; \boldsymbol{\varphi}, V_1, V_2)$, find

$$\mathbb{E}_{V_1, V_2, \boldsymbol{\varphi}} [\mathbb{E}_{\boldsymbol{\theta}} [G_j^2(\boldsymbol{\theta}; \boldsymbol{\varphi}, V_1, V_2)]] = \mathbb{E}_{V_1, V_2, \boldsymbol{\varphi}} [\mathbb{E}_{\boldsymbol{\theta}} [G_k^2(\boldsymbol{\theta}; \boldsymbol{\varphi}, V_1, V_2)]] \quad (62)$$

for all j and k . Thus, averaged across this ensemble $(\boldsymbol{\varphi}, V_1, V_2)$, all loss function parameters belong to the same trainability class. In combination with the preceding variance bound, a polynomially large loss variance implies that the common gradient second moment cannot be exponentially small, provided that the number of trainable parameters grows at most polynomially. Consequently, we can expect that barren plateaus can be ruled out without explicitly evaluating the individual gradients.

C Average linear optical network output

In this section, we compute the average output state from a linear optical network using Weingarten calculus. Alternative proofs based around Schur’s lemma also exist, such as in Ref. [51]. Consider an arbitrary M -mode state ρ with photon number representation

$$\rho = \sum_{\mathbf{n}, \mathbf{m} \in \mathbb{N}_0^M} c_{\rho}(\mathbf{n}, \mathbf{m}) |\mathbf{n}\rangle\langle\mathbf{m}|. \quad (63)$$

We evolve it via a passive linear optical network U . In Theorem 3, we compute the average output state over the Haar measure of $U(M)$.

Theorem 3. *For an M -mode linear optical network U , the Haar averaged output state of $U\rho U^\dagger$ is given by*

$$\mathbb{E}_U [U\rho U^\dagger] = \sum_{n=0}^{\infty} p_{\rho}(n) \pi_M(n) \quad (64)$$

where $\pi_M(n)$ is the maximally mixed state consisting of n photons in M modes, i.e., the normalised projector onto the n photon subspace,

$$\pi_M(n) = \frac{1}{\binom{n+M-1}{n}} \Pi_M(n), \quad \Pi_M(n) = \sum_{\mathbf{m} \in \Phi_M^n} |\mathbf{m}\rangle\langle\mathbf{m}|, \quad (65)$$

and $p_{\rho}(n)$ is the trace of the projector onto the n photon subspace of ρ ,

$$p_{\rho}(n) = \text{tr}[\Pi_M(n) \rho]. \quad (66)$$

Note that this derivation equally applies to arbitrary operators $\hat{O} = \sum_{\mathbf{n}, \mathbf{m}} c_{\hat{O}}(\mathbf{n}, \mathbf{m}) |\mathbf{n}\rangle\langle\mathbf{m}|$.

Proof. The M -mode input state can be written as

$$\rho = \sum_{n,m=0}^{\infty} \sum_{\mathbf{n} \in \Phi_M^n} \sum_{\mathbf{m} \in \Phi_M^m} \frac{c(\mathbf{n}, \mathbf{m})}{\sqrt{n_1! \dots n_M! m_1! \dots m_M!}} \hat{a}_1^{\dagger n_1} \dots \hat{a}_M^{\dagger n_M} |0\rangle\langle 0| \hat{a}_1^{m_1} \dots \hat{a}_M^{m_M} \quad (67)$$

A passive linear optical network U transforms $\hat{a}_j^{\dagger} \rightarrow U_{jk} \hat{a}_k^{\dagger}$ and $\hat{a}_j \rightarrow U_{jk}^* \hat{a}_k$ for the unitary matrix $U \in \text{U}(M)$. Note that, unless otherwise stated, we are using Einstein summation convention where double indices are summed over. The evolved state thus reads

$$\begin{aligned} U\rho U^{\dagger} &= \sum_{n,m=0}^{\infty} \sum_{\mathbf{n} \in \Phi_M^n} \sum_{\mathbf{m} \in \Phi_M^m} \frac{c(\mathbf{n}, \mathbf{m})}{\sqrt{n_1! \dots n_M! m_1! \dots m_M!}} \\ &\quad \times (\delta_{1j} U_{jk} \hat{a}_k^{\dagger})^{n_1} \dots (\delta_{Mj} U_{jk} \hat{a}_k^{\dagger})^{n_M} |0\rangle\langle 0| (\delta_{1j'} U_{j'k'}^* \hat{a}_{k'})^{m_1} \dots (\delta_{Mj'} U_{j'k'}^* \hat{a}_{k'})^{m_M} \\ &= \sum_{n,m=0}^{\infty} \sum_{\mathbf{n} \in \Phi_M^n} \sum_{\mathbf{m} \in \Phi_M^m} \frac{c(\mathbf{n}, \mathbf{m})}{\sqrt{n_1! \dots n_M! m_1! \dots m_M!}} \\ &\quad \times \delta_{1j_1} \dots \delta_{1j_{n_1}} \delta_{2j_{n_1+1}} \dots \delta_{2j_{n_1+n_2}} \delta_{3j_{n_1+n_2+1}} \dots \delta_{Mj_n} \\ &\quad \times \delta_{1j'_1} \dots \delta_{1j'_{m_1}} \delta_{2j'_{m_1+1}} \dots \delta_{2j'_{m_1+m_2}} \delta_{3j'_{m_1+m_2+1}} \dots \delta_{Mj'_m} \\ &\quad \times U_{j_1 k_1} U_{j_2 k_2} \dots U_{j_n k'_n} \times U_{j'_1 k'_1}^* U_{j'_2 k'_2}^* \dots U_{j'_m k'_m}^* \\ &\quad \times \hat{a}_{k'_1}^{\dagger} \hat{a}_{k'_2}^{\dagger} \dots \hat{a}_{k'_n}^{\dagger} |0\rangle\langle 0| \hat{a}_{k'_1} \hat{a}_{k'_2} \dots \hat{a}_{k'_m} \end{aligned} \quad (68)$$

The j indices and Kronecker deltas label which mode the photons came from while the k indices label which mode the photon ends up in. Weingarten calculus [73] can be used to average over $\text{U}(M)$ via

$$\begin{aligned} \int_{\text{U}(M)} d\mu(U) U_{j_1 k_1} \dots U_{j_n k'_n} U_{j'_1 k'_1}^* \dots U_{j'_m k'_m}^* \\ = \delta_{n,m} \sum_{\sigma, \tau \in \mathcal{S}_n} \text{Wg}_{M,n}(\tau \sigma^{-1}) \delta_{j_1 j'_{\sigma(1)}} \dots \delta_{j_n j'_{\sigma(n)}} \delta_{k_1 k'_{\tau(1)}} \dots \delta_{k_n k'_{\tau(n)}} \end{aligned} \quad (69)$$

where $\mathcal{S}_n$ is the symmetric group on n elements and $\text{Wg}_{M,n}$ is the Weingarten function. Plugging this in to Eq. (68) yields the average output state,

$$\begin{aligned} \mathbb{E}_{\mu} [U\rho U^{\dagger}] &= \sum_{n=0}^{\infty} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \frac{c(\mathbf{n}, \mathbf{m})}{\sqrt{n_1! \dots n_M! m_1! \dots m_M!}} \sum_{\sigma, \tau \in \mathcal{S}_n} \text{Wg}_{M,n}(\tau \sigma^{-1}) \\ &\quad \times \delta_{1j_{\sigma(1)}} \dots \delta_{1j_{\sigma(n_1)}} \delta_{2j_{\sigma(n_1+1)}} \dots \delta_{2j_{\sigma(n_1+n_2)}} \dots \delta_{Mj_{\sigma(n)}} \\ &\quad \times \delta_{1j_1} \dots \delta_{1j_{m_1}} \delta_{2j_{m_1+1}} \dots \delta_{2j_{m_1+m_2}} \dots \delta_{Mj_n} \\ &\quad \times \hat{a}_{k'_{\tau(1)}}^{\dagger} \dots \hat{a}_{k'_{\tau(n)}}^{\dagger} |0\rangle\langle 0| \hat{a}_{k_1} \dots \hat{a}_{k_n} \end{aligned} \quad (70)$$

The final line can be computed by noting that $[\hat{a}_k^{\dagger}, \hat{a}_{k'}^{\dagger}] = 0$ and hence τ can be dropped from the creation operator labels; the residual sum over Weingarten functions was com-

puted in Ref [74]

$$\begin{aligned} \sum_{\tau \in \mathcal{S}_n} W_{g_{M,n}}(\tau \sigma^{-1}) &= \sum_{\tau' \in \mathcal{S}_n} W_{g_{M,n}}(\tau') \\ &= \frac{1}{n!} \binom{n+M-1}{n}^{-1}. \end{aligned} \quad (71)$$

Similarly, we can contract the k indices over the creation/annihilation operators to give the projector onto the n photon subspace across M modes multiplied by $n!$; this was done in Ref. [75]

$$\hat{a}_{k_1}^\dagger \dots \hat{a}_{k_n}^\dagger |0\rangle\langle 0| \hat{a}_{k_1} \dots \hat{a}_{k_n} = n! \Pi_M(n). \quad (72)$$

Incorporating these into Eq. (70) gives

$$\begin{aligned} \mathbb{E}_\mu [U \rho U^\dagger] &= \sum_{n=0}^{\infty} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \frac{c(\mathbf{n}, \mathbf{m})}{\sqrt{n_1! \dots n_M! m_1! \dots m_M!}} \binom{n+M-1}{n}^{-1} \Pi_M(n) \\ &\quad \times \sum_{\sigma \in \mathcal{S}_n} \delta_{1j_{\sigma(1)}} \dots \delta_{j_{\sigma(n_1)}} \delta_{2j_{\sigma(n_1+1)}} \dots \delta_{2j_{\sigma(n_1+n_2)}} \dots \delta_{Mj_{\sigma(n)}} \end{aligned} \quad (73)$$

The summand on the second line is only non-zero when $\mathbf{n} = \mathbf{m}$. Further, it is only non-zero when σ is an element of the Young subgroup $Y(\mathbf{n}) = \mathcal{S}_{\{1, \dots, n_1\}} \times \mathcal{S}_{\{n_1+1, \dots, n_1+n_2\}} \times \dots \times \mathcal{S}_{\{n-n_M, \dots, n\}}$, where $\mathcal{S}_{\{\bullet\}}$ is the symmetric group acting on the set $\{\bullet\}$ (i.e., so $\mathcal{S}_n = \mathcal{S}_{\{1, 2, \dots, n\}}$). The order of this Young subgroup is clearly given by $n_1! n_2! \dots n_M!$. It follows that

$$\frac{1}{\sqrt{n_1! \dots n_M! m_1! \dots m_M!}} \sum_{\sigma \in \mathcal{S}_n} \delta_{1j_{\sigma(1)}} \dots \delta_{j_{\sigma(n_1)}} \delta_{2j_{\sigma(n_1+1)}} \dots \delta_{2j_{\sigma(n_1+n_2)}} \dots \delta_{Mj_{\sigma(n)}} = 1 \quad (74)$$

and hence we derive the average output,

$$\begin{aligned} \mathbb{E}_\mu [U \rho U^\dagger] &= \sum_{n=0}^{\infty} \sum_{\mathbf{n} \in \Phi_M^n} c(\mathbf{n}, \mathbf{n}) \binom{n+M-1}{n}^{-1} \Pi_M(n) \\ &= \sum_{n=0}^{\infty} p_\rho(n) \pi_M(n). \end{aligned} \quad (75)$$

□

D First moment of the loss function

In this section, we compute the first moment of the loss function for different initial states. The observables under consideration are constructed from photon number monomials (PNMs), defined as

$$\hat{M}_\alpha = \bigotimes_{i=1}^K \hat{n}_i^{\alpha_i} \quad (76)$$

up to relabelling of the modes, where $\alpha = (\alpha_1, \dots, \alpha_K)$. The k^{th} moment of the PNMs are given by

$$\mu_k(\alpha; \rho) = \mathbb{E}_U \left[\text{tr} \left[\hat{M}_\alpha U \rho U^\dagger \right] \right]. \quad (77)$$

We also consider photon number polynomials (PNPs), given by linear combinations of PNMs; these can be written as

$$\hat{O} = \sum_{\alpha \in A} c_\alpha \hat{M}_\alpha \quad (78)$$

for some $c_\alpha \in \mathbb{R}$ and set of powers A . The set of PNPs are universal for any observable diagonal in the number basis. The first moment of any PNP is just a linear combination of PNM first moments;

$$\mu_1(\hat{O}) = \sum_{\alpha \in A} c_\alpha \mu_1(\alpha) \quad (79)$$

and hence we just need to compute $\mu_1(\alpha)$ for different initial states.

D.1 Derivation for PNMs

Let us consider an input Fock state $|\mathbf{n}\rangle = |n_1, \dots, n_M\rangle$ with total photon number $\sum_{i=1}^M n_i = n$. From Theorem 3, the average output state of $U|\mathbf{n}\rangle$ is just the maximally mixed state of n photons, $\pi_M(n)$. The first moment $\mu_1(\alpha)$ therefore reads

$$\mu_1(\alpha; n) = \binom{n+M-1}{n}^{-1} \sum_{\mathbf{m} \in \Phi_M^n} \left(\prod_{i=1}^M m_i^{\alpha_i} \right) \quad (80)$$

where $\alpha = (\alpha_1, \dots, \alpha_M)$ and can now include zeros (i.e., $\alpha_j = 0$ for $M - K$ entries).

To compute this quantity, we rewrite $m^{\alpha_i} = \sum_{p=0}^{\alpha_i} p! \binom{m_i}{p} S(\alpha_i, p)$ where $S(\alpha_i, p)$ are the Stirling numbers of the second kind.

$$\begin{aligned} \mu_1(\alpha; n) &= \binom{n+M-1}{n}^{-1} \sum_{\mathbf{m} \in \Phi_M^n} \sum_{\mathbf{p}=0}^{\alpha} \left(\prod_{i=1}^M \binom{m_i}{p_i} p_i! S(\alpha_i, p_i) \right) \\ &= \binom{n+M-1}{n}^{-1} \sum_{\mathbf{p}=0}^{\alpha} \left(\prod_{i=1}^M p_i! S(\alpha_i, p_i) \right) \sum_{\mathbf{m} \in \Phi_M^n} \binom{m_i}{p_i} \end{aligned} \quad (81)$$

where $\sum_{\mathbf{p}=0}^{\alpha} = \sum_{p_1=0}^{\alpha_1} \dots \sum_{p_M=0}^{\alpha_M}$. The sums over $\mathbf{m}$ can be expressed using the generating function

$$\begin{aligned} \mathcal{G}(x, p_i) &= \sum_{m_i=0}^{\infty} \binom{m_i}{p_i} x^{m_i} \\ &= \frac{x^{p_i}}{(1-x)^{p_i+1}} \end{aligned} \quad (82)$$

via

$$\begin{aligned}
\sum_{\mathbf{m} \in \Phi_M^n} \left(\prod_{i=1}^M \binom{m_i}{k_i} \right) &= [x^n] \left(\prod_{i=1}^M \mathcal{G}(x, p_i) \right) \\
&= [x^n] \left(x^P (1-x)^{-(P+M)} \right) \\
&= [x^n] \left(\sum_{j=0}^{\infty} (-1)^j \binom{-(P+M)}{j} x^{j+P} \right) \\
&= (-1)^{n-P} \binom{-(P+M)}{n-P} \\
&= \binom{n+M-1}{n-P}
\end{aligned} \tag{83}$$

for $P = \sum_{i=1}^M p_i$. We can feed this into Eq. (81) to give

$$\mu_1(\boldsymbol{\alpha}; n) = \sum_{\mathbf{p}=0}^{\boldsymbol{\alpha}} \frac{\binom{n+M-1}{n-P}}{\binom{n+M-1}{n}} \left(\prod_{i=1}^M p_i! S(\alpha_i, p_i) \right). \tag{84}$$

Noting that if $\alpha_i = 0$ then there is no contribution to the above sum, we can replace $\alpha = (\alpha_1, \dots, \alpha_M)$ for $\alpha_i \geq 0$ with $\alpha = (\alpha_1, \dots, \alpha_K)$ for $\alpha_i > 0$. Then we can carry out each sub-sum associated with a given value of P individually. Note that if $P < K$ then there is at least one $S(\alpha_i, p_i) = 0$, and hence the sum starts at $P = K$. Using this and the convention that $S(\alpha_i, x) = 0$ if $x > \alpha_i$, the first moment can be written as

$$\mu_1(\boldsymbol{\alpha}; n) = \sum_{P=K}^{|\boldsymbol{\alpha}|} \frac{\binom{n+M-1}{n-P}}{\binom{n+M-1}{n}} \sum_{\mathbf{p} \in \Phi_K^P} \left(\prod_{i=1}^K p_i! S(\alpha_i, p_i) \right). \tag{85}$$

This final sum over the $\mathbf{p}$ indices does not admit a simple closed-form structure, but can be computed efficiently using the generating function

$$\mathcal{F}(x, \alpha_i) = \sum_{p_i=0}^{\infty} p_i! S(\alpha_i, p_i) x^{p_i}. \tag{86}$$

Extracting coefficients from a product of generating functions is equivalent to computing the discrete convolution between the analogous sequences.

$$\begin{aligned}
\sum_{\mathbf{p} \in \Phi_K^P} \left(\prod_{i=1}^K p_i! S(\alpha_i, p_i) \right) &= [x^P] \left(\prod_{i=1}^K \mathcal{F}(x, \alpha_i) \right) \\
&= (\mathcal{F}^{\alpha_1} * \mathcal{F}^{\alpha_2} * \dots * \mathcal{F}^{\alpha_K})(P)
\end{aligned} \tag{87}$$

where

$$\mathcal{F}^{\alpha_i} = (j! S(\alpha_i, j))_{j=0}^{\alpha_i} \tag{88}$$

is the sequence associated with generating function $\mathcal{F}(x, \alpha_i)$ and $(f * g)(k)$ is the discrete

convolution between sequences f and g evaluated on the k^{th} index. It follows that the first moment of the cost function for an observable $\hat{M}_{\alpha}$ reads

$$\mu_1(\alpha; n) = \sum_{P=K}^{|\alpha|} \frac{\binom{n+M-1}{n-P}}{\binom{n+M-1}{n}} (\mathcal{F}^{\alpha_1} * \mathcal{F}^{\alpha_2} * \dots * \mathcal{F}^{\alpha_K})(P) \quad (89)$$

Furthermore, from Theorem 3, this can be extended to arbitrary input states; for an initial state ρ with a probability of measuring n photons $p_{\rho}(n)$, the average cost function value is

$$\mu_1(\alpha; \rho) = \sum_{n=K}^{\infty} p_{\rho}(n) \sum_{P=K}^{|\alpha|} \frac{\binom{n+M-1}{n-P}}{\binom{n+M-1}{n}} (\mathcal{F}^{\alpha_1} * \mathcal{F}^{\alpha_2} * \dots * \mathcal{F}^{\alpha_K})(P). \quad (90)$$

Note again that if $n < K$ then the summand vanishes, and hence the summation for n also starts at K .

D.2 Independence from system size

In this section, we study the behaviour of $\mu_1(\alpha)$ when α is held constant but M and n are allowed to grow. We will show that the first moment of the cost function for photon number monomials $\hat{M}_{\alpha}$ asymptotically approaches a constant value if the photon density is held constant with $n = \nu M$ for $\nu \in (0, \infty)$.

Let $\alpha = (\alpha_1, \dots, \alpha_K)$ be a constant such that $K, |\alpha| \ll n, M$. From Eq. (89), the only dependence on n and M is in the binomial terms $\binom{n+M-1}{n-P} / \binom{n+M-1}{n} = \frac{n!}{(n-P)!} \frac{(M-1)!}{(M+P-1)!}$. These scale asymptotically as

$$\frac{n!}{(n-P)!} = n^P \left(1 - \frac{1}{n} \frac{P(P-1)}{2} + \mathcal{O}(n^{-2}) \right) \quad (91)$$

and

$$\frac{(M-1)!}{(M+P-1)!} = M^{-P} \left(1 - \frac{1}{M} \frac{P(P-1)}{2} + \mathcal{O}(M^{-2}) \right). \quad (92)$$

The first moment of the cost function for large n and M is therefore given by

$$\begin{aligned} \mu_1(\alpha; n) &= \sum_{P=K}^{|\alpha|} \left(\frac{n}{M} \right)^P \left(1 - \frac{P(P-1)}{2} (n^{-1} + M^{-1}) \right) (\mathcal{F}^{\alpha_1} * \dots * \mathcal{F}^{\alpha_K})(P) \\ &\quad + \mathcal{O}(1/n^2, 1/M^2, 1/nM) \end{aligned} \quad (93)$$

If the photon number scales with system size as $n = \nu M$, then the first moment asymptotically approaches a constant dependent only on the local photon density, ν , and local observable, $\alpha = (\alpha_1, \dots, \alpha_K)$,

$$\mu_1(\alpha; n = \nu M) = \sum_{P=K}^{|\alpha|} \nu^P (\mathcal{F}^{\alpha_1} * \mathcal{F}^{\alpha_2} * \dots * \mathcal{F}^{\alpha_K})(P) - \mathcal{O}(1/M). \quad (94)$$

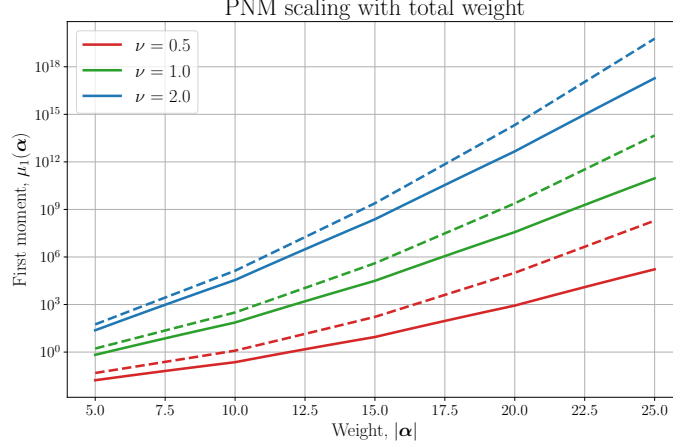


Figure 10: Plot of a PNM first moment against the total weight $|\alpha|$. The solid curve is Eq. (9) at $M = 50$ and the dotted curve is the $M \rightarrow \infty$ limit from Eq. (10). The support is fixed at $K = 5$, and the weight is maximally distributed to be as close to $\alpha_j = |\alpha|/K$ for all j as possible. We see a superexponential increase in the first moment as the total weight of the PNMs increase, as predicted in Eq. (11) and Eq. (96). Increasing the density also leads to order of magnitude changes in the first moment. These results indicate that the sample complexity, $\text{Var}_{\text{samp}}(\alpha) \approx \mu_1(2\alpha)$, for estimating the loss function grows with the order of the largest constituent PNM.

The leading term can be expressed as the product of K ordered Bell polynomials, also known as Fubini polynomials.

$$\lim_{M \rightarrow \infty} \mu_1(\alpha; n = \nu M) = \prod_{j=1}^K F_{\alpha_j}(\nu), \quad F_{\alpha_j}(\nu) = \sum_{p=0}^{\alpha_j} p! S(\alpha_j, p) \nu^p \quad (95)$$

The ordered Bell polynomial asymptotics as α_j grows were derived in Ref. [76] to be

$$F_{\alpha_j}(\nu) \sim \frac{\alpha_j!}{(1 + \nu)[\ln(1 + 1/\nu)]^{\alpha_j + 1}} \quad (96)$$

with corrections exponentially small in α_j . Combining Eq. (95) with Eq. (96), the first moment at constant photon density grows factorially in the order of the photon number operator in each mode α_j for $j \in \{1, \dots, K\}$, and exponentially in the number of supported modes K .

The fact the first moment approaches a constant also implies that to estimate $\ell(\theta)$ for observable $\hat{M}_\alpha$ within a given precision $\varepsilon_{\text{additive}}$, the expected number of shots required by Chebyshev's inequality is independent of system size. This further indicates that the *second* moment of the cost function approaches a constant, via

$$\mu_1(\alpha)^2 \leq \mu_2(\alpha) \leq \mu_1(2\alpha) \quad (97)$$

where $2\alpha = (2\alpha_1, 2\alpha_2, \dots, 2\alpha_K)$. Since both the lower and upper bounds are independent of M , this implies $\mu_2(\alpha)$ is also asymptotically bounded.

D.3 Implications for PNP estimation

Consider an observable $\hat{O} = \sum_{\alpha \in A} c_{\alpha} \hat{M}_{\alpha}$ for PNMs of up to a constant size (i.e., $|\alpha|, K_{\alpha}$ are independent of n or M for all $\alpha \in A$). The expected sample variance for estimating $\ell(\theta)$ is upper bounded by

$$\begin{aligned} \text{Var}_{\text{samp}}(\hat{O}) &\leq \mu_1(\hat{O}^2) \\ &= \sum_{\alpha, \beta \in A} c_{\alpha} c_{\beta} \mu_1(\alpha + \beta) \end{aligned} \quad (98)$$

and hence the expected number of shots required to estimate $\ell(\theta)$ within an accuracy $\varepsilon_{\text{additive}}$ with probability $1 - \delta$ satisfies

$$N \leq \frac{\mu_1(\hat{O}^2)}{\delta \varepsilon_{\text{additive}}^2}. \quad (99)$$

In cases where all $c_{\alpha} > 0$ and $|\alpha|$ are upper bounded by a constant such that $\mu_1(\alpha + \beta) \sim O(1)$, the number of shots required scales as $N \sim O(|A|^2)$. However, if we allow coefficients to be negative then cancellations take place that mean $\mu_1(\hat{O}^2)$ grows more slowly. For example, if we let c_{α} be stochastically selected from a distribution satisfying $\mathbb{E}[c_{\alpha}] = 0$ and $|c_{\alpha}| \sim O(1)$ for all α , then the expected value of $\mu_1(\hat{O}^2)$ over the coefficients is given by

$$\begin{aligned} \mathbb{E}_{\{c_{\alpha}\}} \mu_1(\hat{O}^2) &= \sum_{\alpha \in A} \mathbb{E}_{\{c_{\alpha}\}} [c_{\alpha}^2] \mu_1(2\alpha) + \sum_{\substack{\alpha, \beta \in A, \\ \alpha \neq \beta}} \mathbb{E}_{\{c_{\alpha}\}} [c_{\alpha}] \mathbb{E}_{\{c_{\beta}\}} [c_{\beta}] \mu_1(\alpha + \beta) \\ &= \sum_{\alpha \in A} \mathbb{E}_{\{c_{\alpha}\}} [c_{\alpha}^2] \mu_1(2\alpha) \\ &\sim O(|A|). \end{aligned} \quad (100)$$

assuming that $|\alpha| \sim O(1)$. We would therefore generally expect off-diagonal terms cancelling if the coefficients are equally likely to be positive as negative, leading to a number of shots scaling as $N \sim O(|A|)$. For a given accuracy $\varepsilon_{\text{additive}}$, the number of samples required by the quantum device therefore asymptotically scales as

$$N \sim \begin{cases} \frac{O(|A|^2)}{\delta \varepsilon_{\text{additive}}^2} & \text{generic coefficients,} \\ \frac{O(|A|)}{\delta \varepsilon_{\text{additive}}^2} & \text{with off-diagonal cancellations} \end{cases} \quad (101)$$

D.4 Comparison with the Lim-Oh algorithm

The Lim-Oh algorithm [52] gives a sampling based method for computing expectation values from passive linear optical circuits when the input state and observable are factorisable. To estimate the expectation value of an observable $\hat{O}$ with additive error $\varepsilon_{\text{additive}}$ and success probability $1 - \delta$, the required number of samples, $N^{(\text{LO})}$, scales as

$$N^{(\text{LO})} \sim O\left(\frac{\|\hat{O}_{\text{eff}}\|_2^2 \log(1/\delta)}{\varepsilon_{\text{additive}}^2}\right) \quad (102)$$

for pure product states, where $\hat{O}_{\text{eff}}$ is the effective observable such that $\hat{O} = \hat{O}_{\text{eff}} \otimes \mathbf{1}$. Generating each sample has a time complexity $O(M^2)$.

To estimate a PNP, one approach would be to use the Lim-Oh algorithm to directly estimate the observable $\hat{O}_{\text{eff}}$. However, for PNPs that span the entire system, this norm term $\|\hat{O}_{\text{eff}}\|_2^2$ grows exponentially, leading to exponential classical overhead. Instead, one can use the Lim-Oh algorithm to estimate the constituent PNM expectation values $\text{tr}[\hat{M}_{\alpha}\rho(\boldsymbol{\theta})]$ and then sum these estimates together. We assume that the constituent PNMs are constant with respect to system size such that $\|\hat{M}_{\alpha,\text{eff}}\|_2^2 \sim O(1)$. One can then either (i) directly estimate every single PNM term and sum them together or (ii) Monte-Carlo sample from $\{\alpha\}$ weighted by $|c_{\alpha}|/\sum_{\alpha \in A} |c_{\alpha}|$.

Considering option (i), to derive the classical complexity of the Lim-Oh algorithm we look at the classical variance associated with estimating $\ell(\boldsymbol{\theta}) = \sum_{\alpha \in A} c_{\alpha} \text{tr}[\hat{M}_{\alpha}\rho(\boldsymbol{\theta})]$; we let $\bar{\ell}$ and $\bar{M}_{\alpha}$ be the classical estimate of ℓ and $\text{tr}[\hat{M}_{\alpha}\rho]$ respectively. The total classical estimation variance reads

$$\text{Var}[\bar{\ell}] = \sum_{\alpha \in A} c_{\alpha}^2 \text{Var}[\bar{M}_{\alpha}] + \sum_{\substack{\alpha, \beta \in A, \\ \alpha \neq \beta}} c_{\alpha} c_{\beta} \text{Cov}[\bar{M}_{\alpha}, \bar{M}_{\beta}] \quad (103)$$

Each $\bar{M}_{\alpha}$ is estimated independently and therefore we assume they have uncorrelated errors – this means all covariances satisfy $\text{Cov}[\bar{M}_{\alpha}, \bar{M}_{\beta}] = 0$ for $\alpha \neq \beta$. From Chebyshev's inequality, to estimate ℓ within additive error $\varepsilon_{\text{additive}}$ with probability $1 - \delta$ requires the variance to satisfy

$$\begin{aligned} \text{Var}[\bar{\ell}] &= \sum_{\alpha \in A} c_{\alpha}^2 \text{Var}[\bar{M}_{\alpha}] \\ &\leq \delta \varepsilon_{\text{additive}}^2 \end{aligned} \quad (104)$$

Let us imagine each estimate $\bar{M}_{\alpha}$ used N_{α} samples from the Lim-Oh algorithm, with each sample labelled $m_{\alpha}^{(x)}$ for $x \in \{1, \dots, N_{\alpha}\}$. The estimator of the PNM reads

$$\bar{M}_{\alpha} = \sum_{x=1}^{N_{\alpha}} m_{\alpha}^{(x)} \quad (105)$$

It is unbiased (so $\bar{M}_{\alpha}|_{N_{\alpha} \rightarrow \infty} = \text{tr}[\hat{M}_{\alpha}\rho(\boldsymbol{\theta})]$) and has a variance

$$\text{Var}[\bar{M}_{\alpha}] = \frac{\text{Var}[m_{\alpha}^{(x)}]}{N_{\alpha}} \quad (106)$$

where $\text{Var}[m_{\alpha}^{(x)}]$ is the variance of sample outcomes from the Lim-Oh algorithm. For PNMs of constant size, we assume this satisfies $\text{Var}[m_{\alpha}^{(x)}] \sim O(1)$.

Let us first assume that $N_{\alpha} = N_{\beta} = N_A$. In this case, Eq. (104) can be written as

$$\frac{1}{N_A} \sum_{\alpha \in A} c_{\alpha}^2 \text{Var}[m_{\alpha}^{(x)}] \leq \delta \varepsilon_{\text{additive}}^2 \implies N_A \geq \frac{O(|A|)}{\delta \varepsilon_{\text{additive}}^2} \quad (107)$$

where we used $\sum_{\alpha \in A} O(1) = O(|A|)$. In this case, *each* PNM estimator would require $O(|A|)/\delta\varepsilon_{\text{additive}}^2$ samples. Since there are $|A|$ such PNMs, the total sample complexity of estimating $\ell(\theta)$ with the Lim-Oh algorithm scales as

$$N^{(\text{LO})} \sim O\left(\frac{|A|^2}{\delta\varepsilon_{\text{additive}}^2}\right) \quad (108)$$

to achieve additive error $\varepsilon_{\text{additive}}$ and a success of at least probability $1 - \delta$. Intuitively, if we want $\bar{\ell}$ to have an additive error $\varepsilon_{\text{additive}}$ from summing $|A|$ terms together, then we need the error of each individual PNM estimate to scale as $\varepsilon_{\text{additive}}/\sqrt{|A|}$. Notice that this is independent of the sign of the coefficients c_{α} since only their squared value appeared.

We now briefly look at option (ii), a Monte-Carlo/importance sampling approach. Rather than estimating $\bar{M}_{\alpha}$ for each $\alpha \in A$, we stochastically select a α from the probability distribution $p_{\alpha} = |c_{\alpha}|/S$, where $S = \sum_{\alpha \in A} |c_{\alpha}|$. At each step, we compute one sample from the Lim-Oh algorithm from the stochastically selected α , until we have a total of $N^{(\text{LO})}$ samples $m^{(x)}$ for $x \in \{1, \dots, N^{(\text{LO})}\}$; each sample has an associated $\alpha^{(x)}$. The unbiased estimator $\bar{\ell}$ is then given by

$$\bar{\ell} = \frac{S}{N^{(\text{LO})}} \sum_{x=1}^{N^{(\text{LO})}} \text{sgn}(\alpha^{(x)}) m^{(x)}. \quad (109)$$

From this, it follows that the variance of the estimator scales as $S^2/N^{(\text{LO})}$; since $S = \sum_{\alpha \in A} |c_{\alpha}| \sim O(|A|)$, the total variance $\text{Var}[\bar{\ell}]$ admits the same scaling with respect to $|A|$, δ and $\varepsilon_{\text{additive}}$ as directly estimating each PNM, namely

$$N^{(\text{LO})} \sim O\left(\frac{|A|^2}{\delta\varepsilon_{\text{additive}}^2}\right) \quad (110)$$

Therefore, the classical complexity using one of the best out-of-the-box. Indeed, these derivations apply to any sampling based approach that can estimate a PNM to a given accuracy (i.e., ε_{α}) with a constant number of samples.

These results combined with those from Eq. (101) suggest that a quadratic speedup over the Lim-Oh algorithm is possible for a wide class of PNPs, namely those with suitably many positive and negative terms. Furthermore, the complexity is determined by the number of terms in the PNP, rather than the total number of modes or photons. This means that the computational complexity can be scaled up much faster than the hardware; for an M mode system and taking $|\alpha| \leq L$, the PNP can have up to $|A| \sim O(M^L)$ terms. Finally, the quadratic speedup is independent of whether the system is efficiently trainable or not, because $\varepsilon_{\text{additive}}^2 = \varepsilon^2 \text{Var}_{\text{circ}}$ – the term that indicates the presence or absence of barren plateaus – remains undetermined.

E Second moment of the loss function

We now proceed to compute the second moment of the loss function. We focus on the case of photon number polynomials given by

$$\hat{C} = \sum_{\alpha \in A} c_{\alpha} \hat{M}_{\alpha} \quad (111)$$

The second moment for an n -photon initial state ρ is then defined as

$$\mu_2(\hat{C}) = \mathbb{E}_U \left[\text{tr} \left[\hat{C} U \rho U^{\dagger} \right] \right] \quad (112)$$

In Ref. [51], the authors derived the second moment of particle-number-preserving observable expectation values for input states with a definite particular number. Their expression is in terms of irreducible representations of the unitary group $U(M)$ indexed by k for $k \in \{0, 1, \dots, n\}$. In particular, the second moment of some expectation value reads

$$\mu_2(\hat{C}) = \sum_{k=0}^n \frac{1}{d_k^{(n)}} \|\rho_k^{(n)}\|_2^2 \|\hat{C}_k^{(n)}\|_2^2 \quad (113)$$

where $\hat{X}_k^{(n)}$ is the operator $\hat{X}$ projected onto the k^{th} irreducible representation, $d_k^{(n)}$ is the associated dimension, and $\|\bullet\|_2^2 = \text{tr}[\bullet^{\dagger}\bullet]$ is the Hilbert-Schmidt norm. The projections onto each irreducible representation can be computed using a lowering map [51]

$$L(\bullet) = \sum_{j=1}^M \hat{a}_j \bullet \hat{a}_j^{\dagger} \quad (114)$$

In particular, Ref. [51] proved that

$$\|\hat{X}_k^{(n)}\|_2^2 = \sum_{\ell=0}^k (-1)^{k-\ell} \frac{2k+M-1}{k+\ell+M-1} \frac{\binom{n-\ell}{k-\ell}}{\binom{n+k+M-1}{n-\ell}} \frac{g_{\ell}(\hat{X})}{(n-\ell)!^2} \quad (115)$$

and

$$g_{\ell}(\hat{X}) = \text{tr} \left[(L^{n-\ell}(\hat{X}))^2 \right] \quad (116)$$

The problem is thus reduced to computing $g_{\ell}(\hat{X})$ for $\hat{X} \in \{\rho, \hat{O}\}$.

E.1 Fock state projected norm

From Ref. [51], for an initial Fock state $|\mathbf{n}\rangle = |n_1, n_2, \dots, n_M\rangle$ we have

$$g_{\ell}(|\mathbf{n}\rangle\langle\mathbf{n}|) = (n-\ell)!^2 \sum_{\mathbf{a} \in \Phi_M^{n-\ell}} \left(\prod_{i=1}^M \binom{n_i}{a_i}^2 \right) \quad (117)$$

where $\binom{n_i}{a_i}$ enforces $a_i \leq n_i$. This can be computed using generating the functions

$$\mathcal{A}(x; n_i) = \sum_{a_i=0}^{\infty} \binom{n_i}{a_i}^2 x^{a_i}. \quad (118)$$

from which it follows that

$$\sum_{\mathbf{a} \in \Phi_{M^{n-\ell}}} \left(\prod_{i=1}^M \binom{n_i}{a_i}^2 \right) = [x^{n-\ell}] \left(\prod_{i=1}^M \mathcal{A}(x; n_i) \right). \quad (119)$$

The coefficient can be efficiently extracted using discrete convolution of sequences $\mathcal{A}_i^{n_i}$

$$\mathcal{A}_i^{n_i} = \left(\binom{n_i}{a}^2 \right)_{a=0}^n \quad (120)$$

and hence

$$g_\ell(|\mathbf{n}\rangle\langle\mathbf{n}|) = (n-\ell)!^2 (\mathcal{A}_1^{n_1} * \mathcal{A}_2^{n_2} * \dots * \mathcal{A}_M^{n_M})(n-\ell). \quad (121)$$

E.2 Entangled state projected norm

We now consider equally weighted superpositions over all basis states.

$$|\psi\rangle = \sum_{\mathbf{n} \in \Phi_M^n} \frac{|\mathbf{n}\rangle}{\sqrt{\binom{n+M-1}{n}}}. \quad (122)$$

While not physically motivated, these are the maximally dispersed pure state containing a definite n photons so we expect quantitatively different behaviour than for pure Fock states. Furthermore, these highly correlated states lie beyond domain of applicability of the Lim-Oh algorithm due to the large-scale global correlations between modes.

Applying the lowering operator p times to $\rho = |\psi\rangle\langle\psi|$ gives

$$L^p(\rho) = \frac{1}{\binom{n+M-1}{n}} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \sum_{\substack{\mathbf{c} \in \Phi_M^p \\ \mathbf{c} \leq \min(\mathbf{n}, \mathbf{m})}} \binom{p}{c_1, \dots, c_M} \prod_{i=1}^M a_i^{c_i} |n_i\rangle \langle m_i| (a_i^\dagger)^{c_i}. \quad (123)$$

Just as in Ref. [51]. The action of the creation/annihilation operators together with the multinomial coefficient simplifies to

$$\binom{p}{c_1, \dots, c_M} \prod_{i=1}^M \sqrt{\frac{n_i!}{(n_i - c_i)!}} \sqrt{\frac{m_i!}{(m_i - c_i)!}} = p! \prod_{i=1}^M \sqrt{\binom{n_i}{c_i} \binom{m_i}{c_i}}, \quad (124)$$

so that,

$$L^p(\rho) = \frac{p!}{\binom{n+M-1}{n}} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \sum_{\mathbf{c} \in \Phi_M^p} \left(\prod_{i=1}^M \sqrt{\binom{n_i}{c_i} \binom{m_i}{c_i}} \right) |\mathbf{n} - \mathbf{c}\rangle\langle\mathbf{m} - \mathbf{c}|. \quad (125)$$

where $c_i \leq \min[n_i, m_i]$ is enforced by the binomial coefficient $\binom{x}{y}$ being zero if $y > x$.

Taking the Hilbert–Schmidt norm of Eq. (125) requires two independent copies of $(\mathbf{n}, \mathbf{m}, \mathbf{c})$,

$$\text{tr} [L^p(\rho)^2] = \left(\frac{p!}{\binom{n+M-1}{n}} \right)^2 \sum_{\mathbf{n}_1, \mathbf{n}_2, \mathbf{m}_1, \mathbf{m}_2, \mathbf{c}_1, \mathbf{c}_2} W_1 W_2 \delta_{\mathbf{m}_1 - \mathbf{c}_1, \mathbf{n}_2 - \mathbf{c}_2} \delta_{\mathbf{m}_2 - \mathbf{c}_2, \mathbf{n}_1 - \mathbf{c}_1}, \quad (126)$$

where $W_1 = \prod_i \sqrt{\binom{n_{1,i}}{c_{1,i}} \binom{m_{1,i}}{c_{1,i}}}$. Resolving the deltas leaves the free indices $\mathbf{n}_1, \mathbf{m}_1, \mathbf{c}_1, \mathbf{c}_2$,

$$\text{tr} [L^p(\rho)^2] = \left(\frac{p!}{\binom{n+M-1}{n}} \right)^2 \sum_{\mathbf{n}_1, \mathbf{m}_1, \mathbf{c}_1, \mathbf{c}_2} W_1 W'_2, \quad (127)$$

with $W'_2 = \prod_j \sqrt{\binom{n_1 - \mathbf{c}_1 + \mathbf{c}_2}{c_{2,j}} \binom{m_1 - \mathbf{c}_1 + \mathbf{c}_2}{c_{2,j}}}$. As before, the sum factorises over modes and can be evaluated with a generating function. Defining

$$\begin{aligned} \mathcal{T}_i(x, y, u, v) = & \sum_{n_{1,i}, m_{1,i}, c_{1,i}, c_{2,i} \geq 0} \sqrt{\binom{n_{1,i}}{c_{1,i}} \binom{m_{1,i}}{c_{1,i}} \binom{n_{1,i} - c_{1,i} + c_{2,i}}{c_{2,i}} \binom{m_{1,i} - c_{1,i} + c_{2,i}}{c_{2,i}}} \\ & \times x^{n_{1,i}} y^{m_{1,i}} u^{c_{1,i}} v^{c_{2,i}}, \end{aligned} \quad (128)$$

and $\mathcal{T}(x, y, u, v) = \prod_{i=1}^M \mathcal{T}_i(x, y, u, v)$, the constraints $\mathbf{n}_1, \mathbf{m}_1 \in \Phi_M^n$ and $\mathbf{c}_1, \mathbf{c}_2 \in \Phi_M^p$ are recovered by extracting $[x^n y^n u^p v^p] \mathcal{T}(x, y, u, v)$, equivalently by the convolution

$$(\mathcal{T}_1 * \cdots * \mathcal{T}_M)(n, n, p, p). \quad (129)$$

Replacing p with $n - \ell$ gives

$$g_\ell(|\psi\rangle\langle\psi|) = \left(\frac{(n - \ell)!}{\binom{n+M-1}{n}} \right)^2 (\mathcal{T}_1 * \cdots * \mathcal{T}_M)(n, n, n - \ell, n - \ell). \quad (130)$$

While $\mathbf{c}_1 \in \Phi_M^p$ is enforced automatically by the binomial coefficients in Eq. (128), the same is not true for $\mathbf{c}_2$, whose total is fixed only through the coefficient extraction itself.

E.3 Photon number polynomial projected norm

To compute $g_\ell(\hat{C}) = \text{tr} [L^{n-\ell}(\hat{C})^2]$ for $\hat{C} = \sum_{\alpha \in A} c_\alpha \hat{M}_\alpha$, we can use the fact that

$$L^{n-\ell}(|\mathbf{n}\rangle\langle\mathbf{n}|) = (n - \ell)! \sum_{\mathbf{a} \in \Phi_M^\ell} \left(\prod_{i=1}^M \binom{n_i}{a_i} \right) |\mathbf{a}\rangle\langle\mathbf{a}| \quad (131)$$

from Ref. [51]. The lowering operator on the PNP then gives

$$\begin{aligned} L^{n-\ell}(\hat{C}) &= \sum_{\alpha \in A} c_{\alpha} L^{n-\ell}(\hat{M}_{\alpha}) \\ &= \sum_{\alpha \in A} \sum_{\mathbf{n} \in \Phi_M^n} c_{\alpha} \left(\prod_{i=1}^M n_i^{\alpha_i} \right) L^{n-\ell}(|\mathbf{n}\rangle\langle\mathbf{n}|) \end{aligned} \quad (132)$$

from which we can derive $g_{\ell}(\hat{C}) = \text{tr} [L^{n-\ell}(\hat{C})^2]$,

$$\begin{aligned} g_{\ell}(\hat{C}) &= (n-\ell)!^2 \sum_{\alpha, \beta \in A} c_{\alpha} c_{\beta} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \sum_{\mathbf{a}, \mathbf{b} \in \Phi_M^{\ell}} \left(\prod_{i=1}^M n_i^{\alpha_i} m_i^{\beta_i} \binom{n_i}{a_i} \binom{m_i}{b_i} \right) |\langle \mathbf{a} | \mathbf{b} \rangle|^2 \\ &= (n-\ell)!^2 \sum_{\alpha, \beta \in A} c_{\alpha} c_{\beta} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \sum_{\mathbf{a} \in \Phi_M^{\ell}} \left(\prod_{i=1}^M n_i^{\alpha_i} \binom{n_i}{a_i} \right) \left(\prod_{j=1}^M m_j^{\beta_j} \binom{m_j}{a_j} \right). \end{aligned} \quad (133)$$

The sums over $\mathbf{n}$ and $\mathbf{m}$ can be computed using generating functions. Let us define

$$\mathcal{G}(x; \alpha_i, \mathbf{a}) = \sum_{n=0}^{\infty} n^{\alpha_i} \binom{n_i}{a_i} x^{n_i}. \quad (134)$$

Then it follows that

$$\sum_{\mathbf{n} \in \Phi_M^n} \left(\prod_{i=1}^M n_i^{\alpha_i} \binom{n_i}{a_i} \right) = [x^n] \prod_{i=1}^M \mathcal{G}(x; \alpha_i, a_i) \quad (135)$$

and hence

$$\begin{aligned} &\sum_{\mathbf{a} \in \Phi_M^{\ell}} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \left(\prod_{i=1}^M n_i^{\alpha_i} m_i^{\beta_i} \binom{n_i}{a_i} \binom{m_i}{a_i} \right) \\ &= [x^n y^n] \sum_{\mathbf{a} \in \Phi_M^{\ell}} \left(\prod_{i=1}^M \mathcal{G}(x; \alpha_i, a_i) \mathcal{G}(y; \beta_i, a_i) \right). \end{aligned} \quad (136)$$

Applying a second generating function

$$\mathcal{F}(x, y, z; \alpha_i, \beta_i) = \sum_{a_i=0}^{\infty} \mathcal{G}(x; \alpha_i, a_i) \mathcal{G}(y; \beta_i, a_i) z^{a_i} \quad (137)$$

to the sum over $\mathbf{a}$ allows us to write

$$\sum_{\mathbf{a} \in \Phi_M^{\ell}} \sum_{\mathbf{n}, \mathbf{m} \in \Phi_M^n} \left(\prod_{i=1}^M n_i^{\alpha_i} m_i^{\beta_i} \binom{n_i}{a_i} \binom{m_i}{a_i} \right) = [x^n y^n z^{\ell}] \left(\prod_{i=1}^M \mathcal{F}(x, y, z; \alpha_i, \beta_i) \right). \quad (138)$$

We can therefore write $g_\ell(\hat{C})$ as

$$g_\ell(\hat{C}) = (n - \ell)!^2 \sum_{\alpha, \beta \in A} c_\alpha c_\beta [x^n y^n z^\ell] \left(\prod_{i=1} \mathcal{F}(x, y, z; \alpha_i, \beta_i) \right) \quad (139)$$

The coefficient extraction can be written as the discrete convolutions of the trivariate sequences

$$\mathcal{F}_i^{\alpha_i, \beta_i} = (p^{\alpha_i} q^{\beta_i} \binom{p}{r} \binom{q}{r})_{p, q, r=0}^n \quad (140)$$

associated with each generating function $\mathcal{F}(x, y, z; \alpha_i, \beta_i)$ via

$$[x^n y^n z^\ell] \left(\prod_{i=1}^M \mathcal{F}(x, y, z; \alpha_i, \beta_i) \right) = (\mathcal{F}_1^{\alpha_1, \beta_1} * \mathcal{F}_2^{\alpha_2, \beta_2} * \dots * \mathcal{F}_M^{\alpha_M, \beta_M})(n, n, \ell). \quad (141)$$

For PNPs, we can therefore efficiently calculate g_ℓ with

$$g_\ell(\hat{C}) = (n - \ell)!^2 \sum_{\alpha, \beta \in A} c_\alpha c_\beta (\mathcal{F}_1^{\alpha_1, \beta_1} * \mathcal{F}_2^{\alpha_2, \beta_2} * \dots * \mathcal{F}_M^{\alpha_M, \beta_M})(n, n, \ell). \quad (142)$$

F Gradient estimation with parameter-shift rules

The gradient-based training considered in this work requires estimating derivatives of the loss with respect to individual circuit parameters. In linear optics with Fock input states this can be done exactly through a generalised parameter-shift rule (GPSR) [68, 69, 77, 78], which we briefly outline here. By considering Fourier decompositions of the loss with respect to a single phase-shifter parameter, the first derivative of the loss with respect to a parameter θ_k can be written as,

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_k} = \sum_{j=1}^{2R} \ell(\boldsymbol{\theta} + \kappa_j \mathbf{e}_k) \frac{(-1)^{j+1}}{4R \sin^2(\kappa_j/2)}, \quad \kappa_j = \frac{(2j-1)\pi}{2R}. \quad (143)$$

The gradient of the loss with respect to any circuit parameter is therefore an exact, fixed linear combination of $2R$ loss values at shifted parameter settings. For arbitrary observables, this holds where $R = n$: the number of photons in the circuit. Unlike finite-difference approximations, replacing each loss value in (143) by its empirical estimate $\bar{\ell}_N$ yields an unbiased estimator of the gradient. The same holds for beam splitter parameters, since a tunable beam splitter decomposes into fixed beam splitters and a tunable phase [68].

For the photon number observables considered in this work, the number of shifted loss values needed to compute the gradient can be reduced to the maximum order of the observable. For loss governed by a photon number polynomial, it is sufficient to take $R = |\boldsymbol{\alpha}|$ in (143), where $|\boldsymbol{\alpha}|$ is the greatest order of a photon number monomial appearing in the PNP. For observables whose degree is fixed independently of system size, as for the fixed-order PNPs studied in this work, the number of loss evaluations per gradient component is therefore constant.

Equation (143) reduces gradient estimation to cost function estimation, which is why the analysis of this work centres on the latter. Following the statistical analysis of [68], replacing each loss value in (143) by its empirical estimate from N samples yields an unbiased gradient estimator $\overline{\partial_{\theta_k} \ell}(\boldsymbol{\theta})$. Assuming that average sample variance is equal at shifted loss values (necessarily true for uniform periodic parameters), the average sample variance of the gradient estimate is given by:

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\text{Var} \left[\overline{\partial_{\theta_k} \ell}(\boldsymbol{\theta}) \right] \right] = \frac{2R^2 + 1}{6} \frac{\text{Var}_{\text{samp}}(\hat{O})}{N}. \quad (144)$$

where we have used $\sum_{j=1}^{2R} (4R \sin^2(\kappa_j/2))^{-2} = (2R^2 + 1)/6$ (we derive this following appendix A4 of [79]).

Equation (144) states that the sample variance controls the error of gradient estimates exactly as it does for the loss, up to a factor depending on the number of shifts. This factor scales at worst polynomially as n^2 and, for the case of fixed-order photon number observables, scales as a constant. By Markov's inequality the same holds at typical parameter settings, exactly as in the proof of theorem 1. Appendix B shows the corresponding statement for circuit variance, that the circuit variance of gradients scales with the loss variance Var_{circ} . Consequently, the sample complexity for estimating gradients to within proportional error inherits, up to polynomial prefactors, the $O(\text{Var}_{\text{samp}}(\hat{O})/\text{Var}_{\text{circ}}(\hat{O}))$ scaling of theorem 1, and the sample and circuit variances remain the main quantities governing the complexity of gradient estimation.

G Output binning observable variance

In this section, we compute the Haar-average moments and variances of observables built from output probabilities: single-outcome projectors, and coarse-grained sums of projectors over a bin of outcomes.

G.1 Single-outcome projectors

We consider an input state of fixed photon number n over M modes, and the projector observable $|\mathbf{s}\rangle\langle\mathbf{s}|$ onto an output pattern $\mathbf{s}$. We calculate the required moments:

$$\mu_1(|\mathbf{s}\rangle\langle\mathbf{s}|) = \mathbb{E}_{\boldsymbol{\theta}} [\mathbb{P}(\mathbf{s}|\boldsymbol{\theta})] = |\mathcal{H}|^{-1}, \quad (145)$$

$$\mu_1(|\mathbf{s}\rangle\langle\mathbf{s}|^2) = \mu_1(|\mathbf{s}\rangle\langle\mathbf{s}|) = |\mathcal{H}|^{-1}, \quad (146)$$

$$\mu_2(|\mathbf{s}\rangle\langle\mathbf{s}|) = P_2(M, n) \mu_1(|\mathbf{s}\rangle\langle\mathbf{s}|) = P_2(M, n) |\mathcal{H}|^{-1}, \quad (147)$$

where the first result comes from all outcomes being equally likely in the Haar-average regime, the second from projectors being idempotent, and $P_2(M, n)$ is the normalised av-

erage collision probability [51]. Taking differences between the terms, we get the variances,

$$\text{Var}_{\text{circ}}(|\mathbf{s}\rangle\langle\mathbf{s}|) = (P_2(M, n) - 1)|\mathcal{H}|^{-2}, \quad (148)$$

$$\text{Var}_{\text{samp}}(|\mathbf{s}\rangle\langle\mathbf{s}|) = |\mathcal{H}|^{-1} (1 - P_2(M, n))|\mathcal{H}|^{-1}. \quad (149)$$

G.2 Coarse-grained binning

To calculate the average circuit variance of output binned observables, we will consider uniformly at random selecting output sets S and average variance over them. We consider a fixed input state of n photons across M modes. The circuit variance of $\hat{O}_S$ decomposes as

$$\text{Var}_{\text{circ}}(\hat{O}_S) = \sum_{\mathbf{s} \in S} \text{Var}_{\mu}(\mathbb{P}(\mathbf{s}|\theta)) + \sum_{\mathbf{s} \neq \mathbf{t} \in S} \text{Cov}_{\mu}(\mathbb{P}(\mathbf{s}|\theta), \mathbb{P}(\mathbf{t}|\theta)). \quad (150)$$

Setting $S = \mathcal{H}$ gives $\hat{O}_{\mathcal{H}} = 1$ fixed, hence $\text{Var}_{\text{circ}}(\hat{O}_{\mathcal{H}}) = 0$, and Eq. (150) gives the identity

$$\sum_{\mathbf{s} \neq \mathbf{t} \in \mathcal{H}} \text{Cov}_{\mu}(\mathbb{P}(\mathbf{s}|\theta), \mathbb{P}(\mathbf{t}|\theta)) = - \sum_{\mathbf{s} \in \mathcal{H}} \text{Var}_{\mu}(\mathbb{P}(\mathbf{s}|\theta)). \quad (151)$$

This equation gives us a relation between the average size of covariance between two output probabilities, and the variance of a single output probability. No statistical equivalence between outcomes is assumed, the moments of $\mathbb{P}(\mathbf{s}|\theta)$ will depend on the collision pattern of $\mathbf{s}$, but only the average relation will be required.

Let S be drawn uniformly over subsets of $\mathcal{H}$ of size $|S| = |\mathcal{H}|/K$. The uniform measure over S and the Haar measure are independent, and we will consider finite sums with bounded summands, so expectations may be exchanged freely by Fubini's theorem. Each outcome $\mathbf{s}$ belongs to S with probability $|S|/|\mathcal{H}|$, and each ordered pair $\mathbf{s} \neq \mathbf{t}$ with probability $|S|(|S| - 1)/(|\mathcal{H}|(|\mathcal{H}| - 1))$, so averaging Eq. (150) over S and applying Eq. (151) gives

$$\begin{aligned} \mathbb{E}_S [\text{Var}_{\text{circ}}(\hat{O}_S)] &= \frac{|S|}{|\mathcal{H}|} \sum_{\mathbf{s}} \text{Var}_{\mu}(\mathbb{P}(\mathbf{s}|\theta)) + \frac{|S|(|S| - 1)}{|\mathcal{H}|(|\mathcal{H}| - 1)} \sum_{\mathbf{s} \neq \mathbf{t}} \text{Cov}_{\mu}(\mathbb{P}(\mathbf{s}|\theta), \mathbb{P}(\mathbf{t}|\theta)) \\ &= \frac{|S|}{|\mathcal{H}|} \left(1 - \frac{|S| - 1}{|\mathcal{H}| - 1}\right) \sum_{\mathbf{s}} \text{Var}_{\mu}(\mathbb{P}(\mathbf{s}|\theta)). \end{aligned} \quad (152)$$

The summed variance is $|\mathcal{H}|$ times the outcome-averaged circuit variance of a single-outcome projector;

$$\sum_{\mathbf{s} \in \mathcal{H}} \text{Var}_{\mu}(\mathbb{P}(\mathbf{s}|\theta)) = |\mathcal{H}| \cdot \text{Var}_{\text{circ}}(|\mathbf{s}\rangle\langle\mathbf{s}|) = \frac{P_2(M, n) - 1}{|\mathcal{H}|}. \quad (153)$$

Substituting into Eq. (152) with $|S| = |\mathcal{H}|/K$ and $|\mathcal{H}| \gg 1$ gives the result listed in Eq. (26), namely

$$\mathbb{E}_S [\text{Var}_{\text{circ}}(\hat{O}_S)] \approx \frac{1}{K} \left(1 - \frac{1}{K}\right) \frac{P_2(M, n) - 1}{|\mathcal{H}|}. \quad (154)$$